(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
29 janvier 2004 (29.01.2004)

PCT

(10) Numéro de publication internationale
WO 2004/010324 A2

(51) Classification internationale des brevets[7] : G06F 17/20

(21) Numéro de la demande internationale :
PCT/CH2003/000490

(22) Date de dépôt international : 18 juillet 2003 (18.07.2003)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
02405626.9 19 juillet 2002 (19.07.2002) EP

(71) Déposant *(pour tous les États désignés sauf US)* : ALBERT-INC. S.A. [CH/CH]; Rue Du Simplon 25, CH-1006 Lausanne (CH).

(72) Inventeur; et

(75) Inventeur/Déposant *(pour US seulement)* : GERMAIN, Nicolas [FR/FR]; 37, rue Jean-Pierre Bredy, F-69100 Villeurbanne (FR).

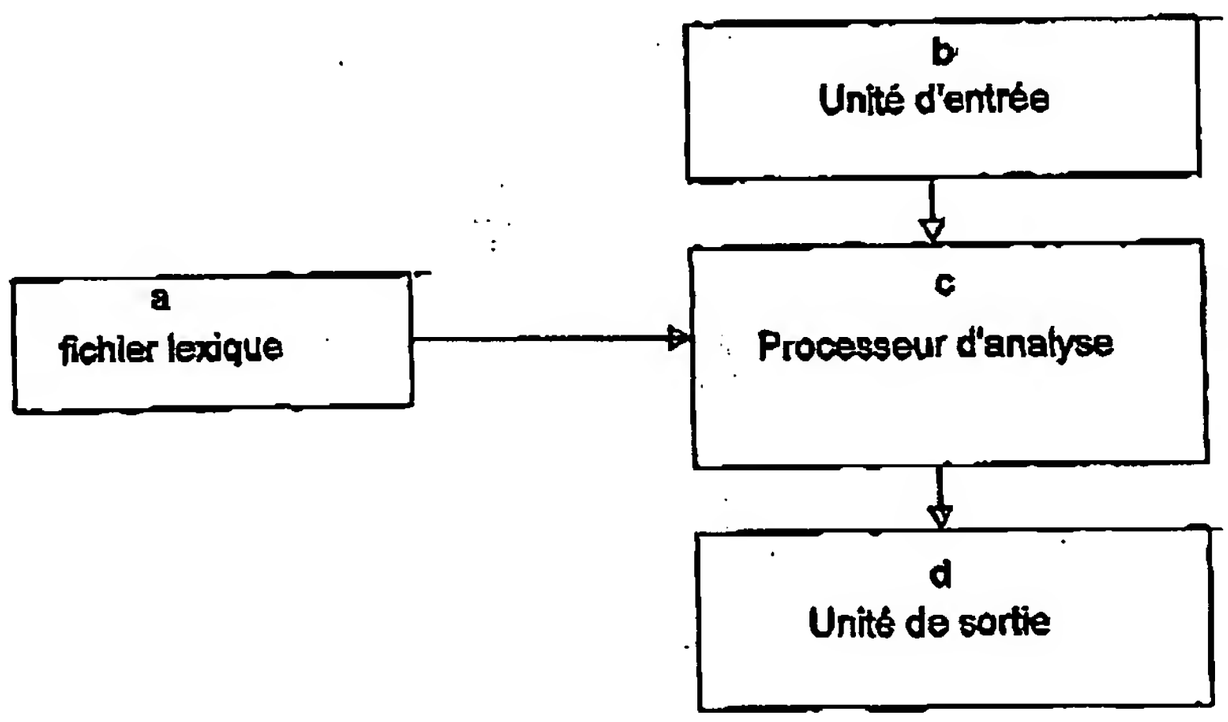
(74) Mandataires : GANGUILLET, Cyril etc.; Abrema Agence Brevets et Marques, Ganguillet & Humphrey, 16, Avenue du Théâtre, Case postale 2065, CH-1002 Lausanne (CH).

(81) États désignés *(national)* : AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Suite sur la page suivante]

(54) Title: SYSTEM FOR EXTRACTING INFORMATION FROM A NATURAL LANGUAGE TEXT

(54) Titre : SYSTEME D'EXTRACTION D'INFORMATIONS DANS UN TEXTE EN LANGAGE NATUREL



a LEXICAL FILE
b INPUT UNIT
c ANALYSIS PROCESSOR
d OUTPUT UNIT

(57) Abstract: The invention relates to a system for extracting information from a natural language text. According to the invention, the extraction method consists in: encoding the words from the text by comparing said words with the contents of a lexicon of empty words (essentially articles, prepositions, conjunctions and verbal auxiliaries); and, subsequently, identifying noun phrases by searching for groups of encoded words that adhere to the pre-defined syntactic rules from among the subsets from the series of encoded words thus obtained.

[Suite sur la page suivante]

WO 2004/010324 A2

WO 2004/010324 A2

(84) **États désignés** *(régional)* : brevet ARIPO (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), brevet eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), brevet européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), brevet OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

— *sans rapport de recherche internationale, sera republiée dès réception de ce rapport*

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

(57) Abrégé : Le procédé d'extraction effectue un codage des mots du texte en les comparant avec le contenu d'un lexique de mots outils (essentiellement articles, prépositions, conjonctions et auxiliaires verbaux), puis identifie des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

WO 2004/010324 PCT/CH2003/000490

Système d'extraction d'informations dans un texte en langage naturel

La présente invention concerne un système d'extraction d'informations dans un texte en langage naturel, en vue de sélectionner les mots ou les groupes de mots du texte qui décrivent le mieux les sujets abordés dans le texte. Ces mots ou groupes de mots sont appelés les "mots-clés" et sont notamment utilisables à des fins d'indexation du texte dans une base de données documentaire, en particulier pour le résumé automatique du texte, pour la catégorisation ou toute autre tentative de représentation de la connaissance.

Les systèmes d'extraction d'informations que l'on connaît et qui tentent d'atteindre ces objectifs utilisent des méthodes d'analyses de trois types :

- les méthodes d'analyse statistique qui tentent d'élire les mots du texte les plus représentatifs en comptant leurs fréquences d'apparition et en ne retenant que ceux dont la fréquence n'est ni trop faible, ni trop forte;

- les méthodes d'analyse à thesaurus qui fonctionnent d'après une représentation prédéfinie de la connaissance et qui sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus. Cette définition est entièrement manuelle et doit être opérée dans chaque domaine de spécialités;

- les méthodes d'analyse à reconnaissance de motifs (patterns) qui fonctionnent à l'aide d'identifications statistiques de motifs (patterns).

Le fonctionnement comparatif de ces trois types de méthodes d'analyse va être illustré ci-après par l'analyse du texte suivant :

"«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays."

(source Reuter)

Fonctionnement des méthodes d'analyse statistique :

Si l'on considère leur approche de façon caricaturale, les méthodes d'analyse statistique comptent les mots du texte pour ne retenir que ceux dont la fréquence n'est ni trop faible ni trop forte en éliminant parfois les mots outils (articles, prépositions, conjonctions, auxiliaires verbaux), afin d'affiner les résultats. En ce qui concerne le texte proposé ci-dessus, les mots "moyennement" fréquents (sans prendre en considération les mots outils) sont alors :

affiche, années, Cats, comédie, dernière, été, longtemps, musicale et représentation.

Bien que le principal avantage des méthodes d'analyse statistique réside dans une grande simplicité algorithmique, leur principal désavantage réside en la faible pertinence des résultats. En effet, les mots "moyennement" fréquents d'un texte sont rarement les plus représentatifs. Ces méthodes peuvent toutefois donner de meilleurs résultats sur des textes plus longs que le texte d'exemple ci-dessus.

D'autre part, du fait que le texte est découpé en mots, c'est-à-dire en chaînes de caractères dont les délimiteurs sont des espaces, les liens sémantiques qui peuvent relier des mots entre eux, comme par exemple les mots "comédie" et "musicale", sont perdus.

Fonctionnement des méthodes d'analyse à thesaurus :

Ces méthodes sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus, cette définition étant, comme on l'a mentionné plus haut, entièrement manuelle et devant être opérée dans chaque domaine de spécialité.

Imaginons par exemple le thesaurus suivant :

spectacle → comédie (s) → dramatique
→ musicale → Cats
→ Les dix commandements
→ savante

Avec ce type de méthodes, il est toujours possible d'identifier les mots du texte source qui se retrouvent exactement sous la même forme dans le thesaurus. L'avantage de ces méthodes est que l'on peut être sûr que les mots identifiés correspondent à une réalité culturelle ou scientifique établie et répertoriée. D'autre part, il est possible de déduire un mot fédérateur comme "spectacle" qui ne fait pas partie du texte initial, mais qui le caractérise correctement. En revanche, l'inconvénient majeur de ces méthodes est qu'il faut perpétuellement mettre à jour le thesaurus pour qu'il conserve sa pertinence, ce qui entraîne des frais de maintenance importants. Un autre inconvénient important de ces méthodes réside dans le fait qu'un thesaurus constitué pour analyser des textes dans le domaine de la chimie ne pourra pas être utilisé pour des textes dans le domaine de l'électronique, par exemple. De plus, dans le cas où le thesaurus n'est pas exhaustif, certaines expressions qui

peuvent être très pertinentes ne seront pas reconnues comme telles.

Fonctionnement des méthodes d'analyse à reconnaissance de motifs :

Les méthodes d'analyse à reconnaissance de motifs que l'on connaît sont des méthodes d'identification statistiques de motifs qui, bien qu'elles améliorent considérablement les méthodes d'analyse statistique mentionnées plus haut, en conservant la trace de l'appariement des mots, comme par exemple des termes "comédie" et "musicale" de l'exemple ci-dessus, ne permettent pas d'analyser de façon correcte des textes courts. En effet, les méthodes statistiques ont besoin de quantité pour fonctionner correctement.

Par exemple, les motifs-clés du texte d'exemple seront obtenus par comparaisons approximatives de séquences plus ou moins longues entre elles. Les mots outils (le, la, les, ...) ne comptent pas, et les séquences sont formées à partir d'un mot, plus ou moins trois mots :

Cats
Cats comédies
Cats comédies musicales
Cats comédies musicales longtemps
comédies
comédies musicales
comédies musicales longtemps
comédies musicales longtemps affiche
musicales
musicales longtemps
musicales longtemps affiche
musicales longtemps affiche tirer
etc_

Il suffit ensuite de regrouper les différentes séquences obtenues, par approximation sur la forme (par exemple

« comédies » et « comédie »), et de compter les expressions combinées les plus fréquentes comme « comédies musicales ».

Le but de la présente invention est de proposer un système pour l'extraction d'informations dans un texte en langage naturel permettant de remédier aux inconvénients des méthodes d'analyses connues, en permettant notamment une analyse de bonne qualité de textes aussi bien courts que longs.

Ce système utilise une méthode d'analyse par identification de motifs (patterns) non pas purement statistique, mais également syntaxique.

En résumé, le système proposé convertit les mots du texte en suite de catégories syntaxiques, puis confronte des sous-ensembles du texte avec des motifs syntaxiques prédéfinis, de façon à identifier des groupes nominaux sans préjuger de la valeur des mots qui composent ces groupes.

Ainsi, les mots « pomme de terre » ou « électronique de puissance » ne sont pas important par eux-mêmes, mais sont importants par rapport au texte où ils apparaissent. Dans un texte de nature générale « électronique de puissance » peut n'être qu'un exemple, pas un mot-clé du texte, mais sera probablement mot-clé dans un texte traitant des transistors. C'est le contexte qui fait le mot-clé, et le système selon la présente invention comporte en quelque sorte un analyseur de contextes syntaxiques. De même, le mot "porte" peut être reconnu comme nominal dans certains textes à cause de sa position par rapport aux autres mots du texte, ou simplement comme mot structurel dans d'autres textes.

Une méthode d'analyse par identification de motifs est proposée dans le document US 4,864,501. Le procédé décrit dans ce document antérieur utilise, pour le codage des mots du texte en vue de l'identification des motifs, un dictionnaire contenant les mots radicaux (base forms). Outre le fait que ce dictionnaire est très volumineux puisqu'il contient plusieurs

dizaines de milliers d'entrées, le procédé nécessite des algorithmes complexes de radicalisation des mots, spécifiques à chaque langue, pour retomber sur des mots du dictionnaire, ainsi qu'éventuellement des tables spécifiques de préfixes/suffixes pour traiter les cas d'erreurs d'orthographe, etc. Il s'agit par conséquent d'un procédé très lourd à mettre en œuvre et à utiliser.

Le système d'extraction selon la présente invention permet de remédier à ces inconvénients.

A cet effet, l'invention concerne un procédé d'extraction d'informations dans un texte en langage naturel, par identification de motifs (patterns), selon lequel on effectue un codage des mots du texte en les comparant avec le contenu d'un lexique prédéfini contenant quelques dizaines de mots outils, et selon lequel on identifie ensuite des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

L'invention concerne également un système d'extraction d'informations dans un texte en langage naturel comprenant :

- une unité d'entrée pour recevoir ledit texte en langage naturel,
- un fichier lexique dans lequel sont enregistrés des mots outils,
- un processeur d'analyse relié à ladite unité d'entrée, au fichier lexique et agencé pour effectuer dans un premier temps le codage des mots dudit texte en langage naturel par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit fichier lexique de mots outils, de façon d'une part à repérer les mots outils dans le texte et à évaluer la fonction des mots d'usage, non reconnus comme mots outils, en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils, et dans un deuxième temps une recherche, parmi

des sous-ensembles de la suite de mots codés obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies, de façon à identifier des groupes nominaux,

- une unité de sortie reliée audit processeur d'analyse pour recevoir les groupes de mots codés reconnus comme des motifs syntaxiques.

Le système d'extraction selon l'invention évalue la fonction grammaticale des mots du texte à analyser à l'aide d'un lexique prédéfini contenant les quelques dizaines de mots outils propres à chaque langue et qui sont essentiellement les articles, les prépositions, les conjonctions et auxiliaires verbaux. La fonction des autres mots est ensuite déduite grâce à l'emplacement des seuls mots outils. Du fait que les mots outils d'un texte représentent couramment 40 à 50 % des mots de ce texte, ceux-ci sont donc toujours assez nombreux pour permettre l'évaluation des autres mots. Ensuite, seules les parties du texte dont la grammaire est identifiée comme mots-clés possibles sont retenues.

Les avantages du système d'extraction selon l'invention sont nombreux. En particulier, le lexique de mots outils utilisé par le système est incomparablement plus léger que les dictionnaires contenant plusieurs milliers de mots qu'utilisent les systèmes connus. On relèvera, d'autre part, qu'aucune intervention humaine n'est nécessaire pour la détermination des mots-clés, que le système peut fonctionner pour des textes de langues diverses et que, mis à part le lexique des mots outils, il ne nécessite aucun autre lexique. De plus, du fait que la valeur sémantique et grammaticale des mots outils est fixe et n'évolue pratiquement jamais sur plusieurs décennies, la maintenance du lexique est des plus réduites. En revanche, la valeur des autres mots, que l'on peut appeler les mots d'usage (verbes, noms, adjectifs), évolue sans cesse dans le temps, en fonction des usages, de l'évolution des métiers ou des sciences, ou simplement en fonction de l'actualité. Du fait que le système de la présente

invention ne présuppose rien sur la valeur des mots d'usage, il fonctionne de façon identique dans tous les domaines, littéraire, technique ou scientifique, alors que les systèmes qui utilisent les méthodes connues doivent toujours être enrichis avec des lexiques spécialisés, fabriqués bien souvent sur mesure. Enfin, ce système d'extraction permet d'adresser de nouvelles langues incomparablement plus rapidement que n'importe quel autre système proposé jusqu'ici.

D'autre part, contrairement aux systèmes utilisant des méthodes d'analyse statistique dans lesquelles la fréquence d'apparition des mots est un critère de sélection, ce qui suppose que le texte soit suffisamment long, le système selon l'invention n'accorde à la fréquence d'apparition des mots qu'une importance subalterne et fonctionne aussi bien pour des textes longs de plusieurs dizaines de pages que pour des textes courts de quelques lignes.

On va décrire ci-après, à titre d'exemple, un système d'extraction d'informations selon l'invention dans un texte en langage naturel, en se référant aux dessins, sur lesquels :

- la fig. 1 est un schéma-bloc du système d'extraction selon l'invention;

- la fig. 2 est un schéma-bloc des étapes d'un mode d'exécution du procédé selon l'invention.

L'utilisation d'un modèle syntaxique requiert de reconnaître la langue du texte analysé. C'est donc naturellement la première opération qu'effectue le système d'extraction selon l'invention. Cette reconnaissance de la langue peut être basée sur des critères purement statistiques de cooccurrence de lettres. La reconnaissance des langues, par exemple anglais, espagnol, français, portugais, allemand ou italien, permet d'orienter les analyses qui seront réalisées en aval.

L'étape suivante est une étape de profilage du texte qui permet d'identifier les lignes de texte (paragraphes) comportant une information linguistique, et d'opérer des regroupements de paragraphes. Cette opération est particulièrement utile pour les textes structurés (avec titres, sous-titres, etc.), car elle permet de regrouper des paragraphes de façon cohérente. Elle est inutile pour des textes courts.

L'étape suivante consiste en une opération de régularisation du texte au cours de laquelle il s'agit d'éliminer les amalgames de signes, comme par exemple séparer les caractères typographiques des caractères alphabétiques. Il sera par exemple utile de reconnaître la chaîne "mot," comme le terme "mot" suivit de ",", alors que la chaîne "1,5" devra être reconnue comme un nombre.

Dans le texte d'exemple, cette étape revient à séparer les caractères typographiques (", " , "'" et ".") des autres mots par des espaces blancs. Le texte d'exemple devient alors :

"« Cats » , l ' une des comédies musicales les plus longtemps à l ' affiche , va tirer sa révérence après vingt et une années sur la scène londonienne . La dernière représentation de cette œuvre d ' Andrew Lloyd Webber aura lieu le 11 mai , jour de son 21e anniversaire , après quelque 9 000 représentations . L ' annonce a été faite trois jours après la dernière représentation de « Starlight Express » , la seconde comédie musicale la plus longtemps à l ' affiche à Londres , après dix-huit années sur les planches .
La fin de « Cats » est un coup dur supplémentaire pour le quartier de Covent Garden , où sont regroupés la plupart des théâtres londoniens , et qui a souffert d ' une forte baisse de fréquentation en 2001 . Depuis 1981 , année de son lancement , la comédie musicale a , depuis , été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays . "

L'étape suivante, qui constitue une étape clé du système, consiste à déterminer la catégorie de chaque mot. Grâce au lexique restreint des mots outils, les mots du texte sont codés selon des catégories grammaticales attribuées en fonction de la valeur syntaxique des mots. Les mots outils du lexique sont dans un premier temps reconnus dans le texte,

puis la fonction des autres mots du texte est déduite en fonction de leur emplacement par rapport aux mots outils déjà reconnus.

Ainsi, si l'on adopte par exemple les catégories suivantes :

s: mot de structure (mot outil non utile pour la suite de l'analyse)
d: déterminant (le, la, les, etc.)
p: préposition (de, en, par, etc.)
4: signe ouvrant ou fermant
1 ou 2 : ponctuation
3: apostrophe
N: nombre
W: nom propre
w: nom commun
c: amalgame (du, des, au, aux, ...)
a: anaphores (ce, cet, ces, ...)
*: code attribué si aucune des catégories précédentes n'est reconnue

Le texte d'exemple mentionné plus haut devient :

4 W 4 2 d 3 d c w3 w4 d w1 w2 p d 3 w3 2 s w2 a w4 w2 w1 p d w2 p d w2 w4 1 d w3 w5 p a w2 p 3 W W W w2 w1 d N w1 2 w1 p a * w5 2 w2 d N N w5 1 d 3 w3 s w2 w2 d w1 w2 d w3 w5 p 4 W W 4 2 d w3 w3 w4 d w1 w2 p d 3 w3 p W 2 w2 d 0 d w2 p d w2 1 d w1 p 4 W 4 s d w1 w1 w5 p d w2 p W W 2 s s w3 d w2 c w2 w3 2 p s s w2 p 3 d w2 w2 p w4 p N 1 W N 2 w2 p a w3 2 d w3 w4 s 2 w2 2 w2 w4 w2 w1 p N w2 p w3 p N w2 p p N w1 1

Une étape suivante consiste à identifier les structures linguistiques appelées syntagmes nominaux dans la terminologie linguistique ou, plus simplement, groupes nominaux.

L'ensemble des motifs syntaxiques qu'il est utile d'identifier constitue la grammaire d'analyse. Du fait que cette grammaire est commune à l'ensemble des langues romanes, il est possible d'analyser un grand nombre de langues en utilisant un même système d'extraction selon l'invention sans adaptation lourde.

A titre d'exemple, une grammaire (simplifiée) peut avoir la forme suivante :

(1) syntagme nominal -> déterminant , groupe nominal ; W .
(2) déterminant -> d ; d , 3 ; nombre ; c ; a
(3) d -> 'le' ; 'la' ; 'les' ; 'des' ; 'l' ; etc…
(3bis) c -> 'du' ; 'au' ; 'aux' ; etc…
(3ter) a -> 'ce' ; 'cette' ; 'ces' ; 'son' ; etc…
(4) groupe nominal -> expression , groupe nominal .
(5) expression -> w , p , w ; w .
(6) p -> 'de' ; 'à' ; 'pour' ; 'sans' ; etc…

La flèche se dit « se réécrit », la virgule se dit « suivi de », le point-virgule exprime un « ou », le point marque la fin de la règle. La règle (1) se lit « syntagme nominal se réécrit déterminant suivi de groupe nominal ».

Les règles (3) et (6) sont dites règles terminales car elles font appels aux formes lexicales du lexique des mots outils.

La règle (4) est une règle récursive. Un groupe nominal peut donc contenir une infinité d'expressions, lesquelles, selon la règle (5) sont soit de type wpw, soit de type w.

Les suites de catégories grammaticales suivantes seront donc reconnues comme syntagme nominal :

d w
d w p w
d w w
d w w p w
d 3 w w
etc…

Sur le texte d'exemple, les groupes nominaux identifiés à l'aide de cette grammaire ont été soulignés :

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelques 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.

(source Reuter)

Comme les groupes nominaux représentent à peu près 50 % du texte, il est nécessaire de ne retenir que ceux dont la probabilité d'être de vrais mots-clés du texte est la plus forte.

Une étape suivante peut consister à filtrer les groupes nominaux. Tous les groupes nominaux n'ont pas la même capacité référentielle. Certains sont plus importants que d'autres. Pour déterminer quels sont les plus importants d'entre eux, le système selon l'invention valorise chaque groupe nominal en fonction d'un double critère, l'un statistique, l'autre syntaxique.

Le critère statistique :

Les mots les plus fréquents des groupes nominaux sont classés par ordre de fréquence décroissant (en tenant compte d'une approximation comme 'comédie' = 'comédies'), soit dans le texte d'exemple :

comédie	3
musicale	3
affiche	2

années	2
Cats	2
dernière	2
représentation	2

Seuls les mots dont l'occurrence dépasse 1 sont conservés dans la liste. Les mots éliminés ont donc une valeur nulle. On ajoute à la valeur de chaque groupe nominal (initialement fixée à 0), la valeur de l'occurrence des mots qu'il contient moins 1. La valeur des groupes nominaux devient :

comédie musicale	(3 - 1) + (3 - 1) = 4
affiche	2 - 1 = 1
affiche à Londres	2 - 1 = 1
Cats	2 - 1 = 1
etc_	

Le critère syntaxique :

Lorsque qu'un groupe nominal est ou comporte un nom propre, celui-ci prend un point de valeur supplémentaire, 0 sinon.

comédie musicale	4 + 0 = 4
affiche	1 + 0 = 1
affiche à Londres	1 + 1 = 2
Cats	1 + 1 = 2
etc_	

Avec cette valorisation, il est aisé de procéder au classement des groupes nominaux. Dans le texte d'exemple, les groupes nominaux perçus comme les plus importants sont soulignés deux fois, les groupes d'importance secondaire sont soulignés une fois, tandis que les autres ont été purement et simplement éliminés.

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le

11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.

(source Reuter)

Revendications

1. Procédé d'extraction d'informations dans un texte en langage naturel, par identification de motifs (patterns), caractérisé en ce que l'on effectue un codage des mots du texte en les comparant avec le contenu d'un lexique prédéfini contenant quelques dizaines de mots outils, et en ce que l'on identifie ensuite des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

2. Procédé selon la revendication 1, caractérisé en ce que le codage des mots du texte s'effectue par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit lexique de mots outils, de façon à repérer les mots outils dans le texte et en ce que la fonction des mots d'usage, non reconnus comme mots outils, est déduite en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils.

3. Procédé selon l'une des revendications 1 ou 2, caractérisé en ce que les groupes nominaux identifiés sont ensuite valorisés de façon à ne retenir que les groupes perçus comme les plus importants en utilisant des critères de valorisation prédéfinis.

4. Système d'extraction d'informations dans un texte en langage naturel, caractérisé en ce qu'il comprend :

- une unité d'entrée pour recevoir ledit texte en langage naturel,
- un fichier lexique dans lequel sont enregistrés des mots outils,
- un processeur d'analyse relié à ladite unité d'entrée, au fichier lexique et agencé pour effectuer dans un premier temps le codage des mots dudit texte en langage naturel par évaluation de la fonction grammaticale de chaque mot en le comparant avec le

contenu dudit fichier lexique de mots outils, de façon d'une part à repérer les mots outils dans le texte et à évaluer la fonction des mots d'usage, non reconnus comme mots outils, en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils, et dans un deuxième temps une recherche, parmi des sous-ensembles de la suite de mots codés obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies, de façon à identifier des groupes nominaux,

- une unité de sortie reliée audit processeur d'analyse pour recevoir les groupes de mots codés reconnus comme des motifs syntaxiques.

5. Système selon la revendication 4, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de valorisation des groupes de mots codés retenus de façon à ne retenir que les groupes perçus comme les plus importants.

6. Système selon l'une des revendications 3 ou 4, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de reconnaissance de la langue du texte reçu dans l'unité d'entrée.

7. Système selon l'une des revendications 4 à 6, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de régularisation du texte reçu dans l'unité d'entrée de façon à éliminer les amalgames de signes.

10/524624

WO 2004/010324

PCT/CH2003/000490

Fig. 1

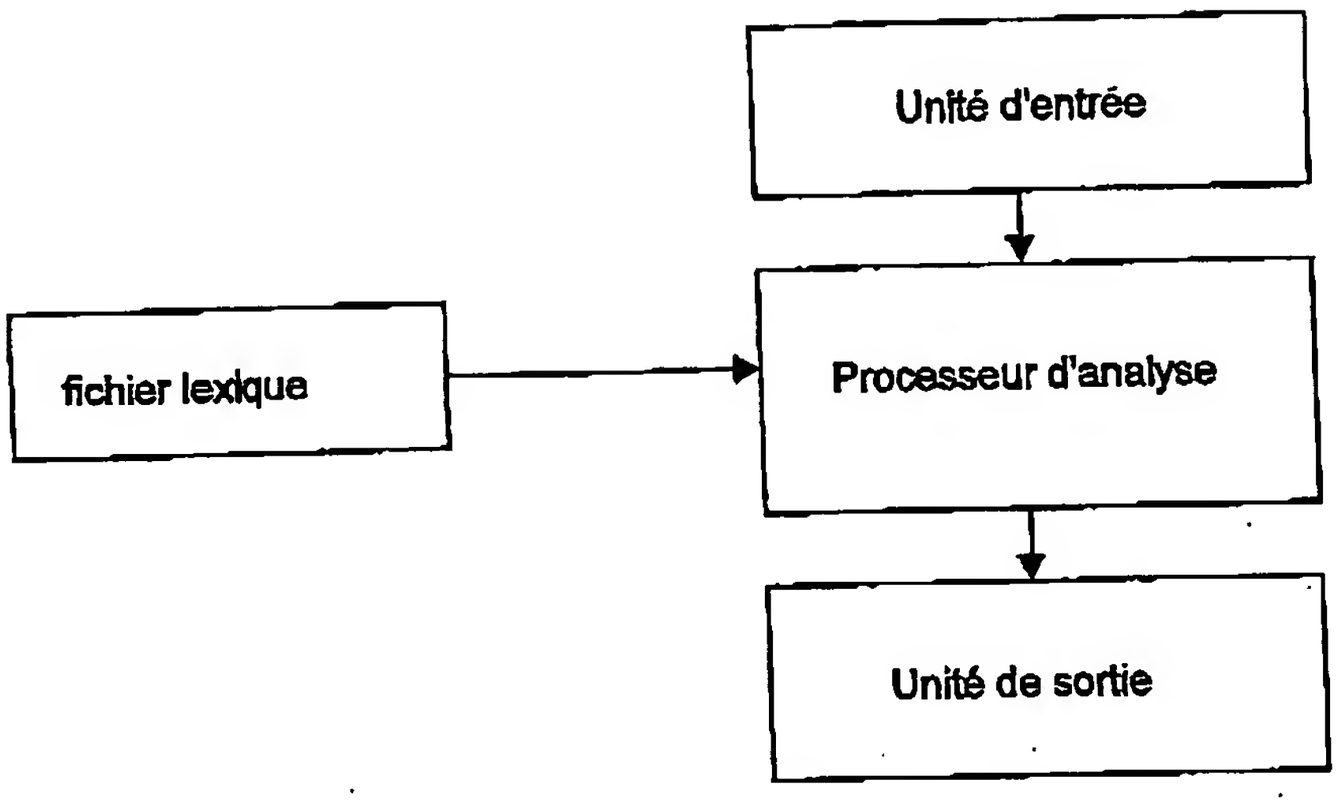
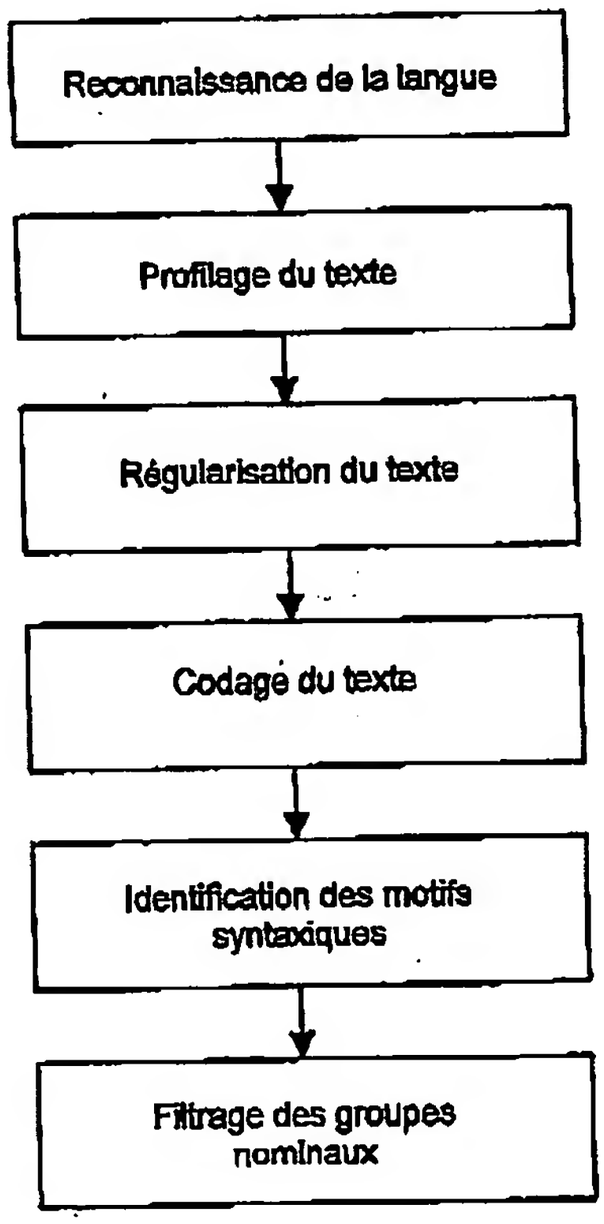


Fig. 2



(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
29 janvier 2004 (29.01.2004)

PCT

(10) Numéro de publication internationale
WO 2004/010324 A2

(51) Classification internationale des brevets[7] : G06F 17/20

(21) Numéro de la demande internationale :
PCT/CH2003/000490

(22) Date de dépôt international : 18 juillet 2003 (18.07.2003)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
02405626.9 19 juillet 2002 (19.07.2002) EP

(71) Déposant *(pour tous les États désignés sauf US)* : **ALBERT-INC. S.A.** [CH/CH]; Rue Du Simplon 25, CH-1006 Lausanne (CH).

(72) Inventeur; et

(75) Inventeur/Déposant *(pour US seulement)* : **GERMAIN, Nicolas** [FR/FR]; 37, rue Jean-Pierre Bredy, F-69100 Villeurbanne (FR).

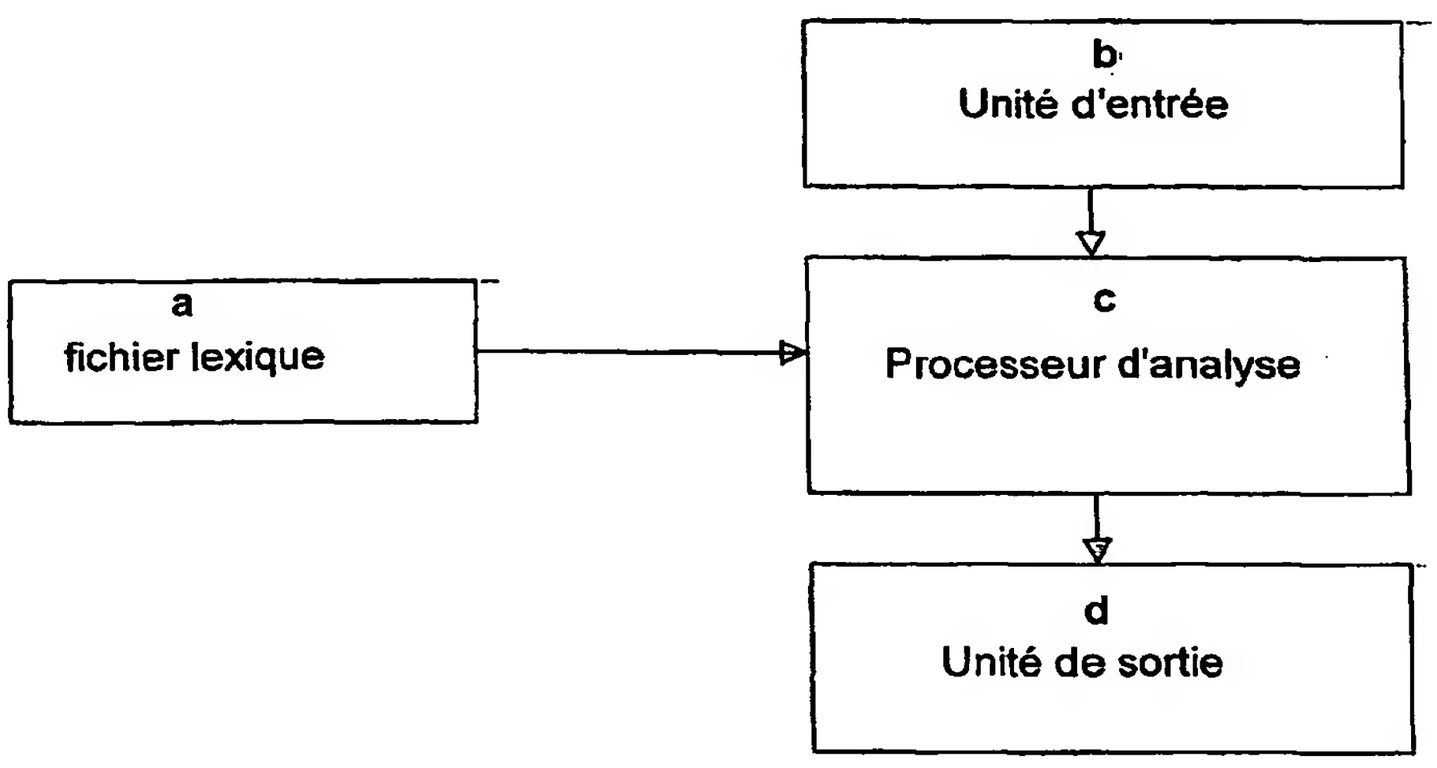
(74) Mandataires : **GANGUILLET, Cyril** etc.; Abrema Agence Brevets et Marques, Ganguillet & Humphrey, 16, Avenue du Théâtre, Case postale 2065, CH-1002 Lausanne (CH).

(81) États désignés *(national)* : AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Suite sur la page suivante]

(54) Title: SYSTEM FOR EXTRACTING INFORMATION FROM A NATURAL LANGUAGE TEXT

(54) Titre : SYSTEME D'EXTRACTION D'INFORMATIONS DANS UN TEXTE EN LANGAGE NATUREL



a LEXICAL FILE
b INPUT UNIT
c ANALYSIS PROCESSOR
d OUTPUT UNIT

(57) **Abstract:** The invention relates to a system for extracting information from a natural language text. According to the invention, the extraction method consists in: encoding the words from the text by comparing said words with the contents of a lexicon of empty words (essentially articles, prepositions, conjunctions and verbal auxiliaries); and, subsequently, identifying noun phrases by searching for groups of encoded words that adhere to the pre-defined syntactic rules from among the subsets from the series of encoded words thus obtained.

[Suite sur la page suivante]

WO 2004/010324 A2

WO 2004/010324 A2

(84) **États désignés** *(régional)* : brevet ARIPO (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), brevet eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), brevet européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), brevet OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

— *sans rapport de recherche internationale, sera republiée dès réception de ce rapport*

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

(57) Abrégé : Le procédé d'extraction effectue un codage des mots du texte en les comparant avec le contenu d'un lexique de mots outils (essentiellement articles, prépositions, conjonctions et auxiliaires verbaux), puis identifie des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

Système d'extraction d'informations dans un texte en langage naturel

La présente invention concerne un système d'extraction d'informations dans un texte en langage naturel, en vue de sélectionner les mots ou les groupes de mots du texte qui décrivent le mieux les sujets abordés dans le texte. Ces mots ou groupes de mots sont appelés les "mots-clés" et sont notamment utilisables à des fins d'indexation du texte dans une base de données documentaire, en particulier pour le résumé automatique du texte, pour la catégorisation ou toute autre tentative de représentation de la connaissance.

Les systèmes d'extraction d'informations que l'on connaît et qui tentent d'atteindre ces objectifs utilisent des méthodes d'analyses de trois types :

- les méthodes d'analyse statistique qui tentent d'élire les mots du texte les plus représentatifs en comptant leurs fréquences d'apparition et en ne retenant que ceux dont la fréquence n'est ni trop faible, ni trop forte;

- les méthodes d'analyse à thesaurus qui fonctionnent d'après une représentation prédéfinie de la connaissance et qui sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus. Cette définition est entièrement manuelle et doit être opérée dans chaque domaine de spécialités;

- les méthodes d'analyse à reconnaissance de motifs (patterns) qui fonctionnent à l'aide d'identifications statistiques de motifs (patterns).

Le fonctionnement comparatif de ces trois types de méthodes d'analyse va être illustré ci-après par l'analyse du texte suivant :

"«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays."

(source Reuter)

Fonctionnement des méthodes d'analyse statistique :

Si l'on considère leur approche de façon caricaturale, les méthodes d'analyse statistique comptent les mots du texte pour ne retenir que ceux dont la fréquence n'est ni trop faible ni trop forte en éliminant parfois les mots outils (articles, prépositions, conjonctions, auxiliaires verbaux), afin d'affiner les résultats. En ce qui concerne le texte proposé ci-dessus, les mots "moyennement" fréquents (sans prendre en considération les mots outils) sont alors :

affiche, années, Cats, comédie, dernière, été, longtemps, musicale et représentation.

Bien que le principal avantage des méthodes d'analyse statistique réside dans une grande simplicité algorithmique, leur principal désavantage réside en la faible pertinence des résultats. En effet, les mots "moyennement" fréquents d'un texte sont rarement les plus représentatifs. Ces méthodes peuvent toutefois donner de meilleurs résultats sur des textes plus longs que le texte d'exemple ci-dessus.

D'autre part, du fait que le texte est découpé en mots, c'est-à-dire en chaînes de caractères dont les délimiteurs sont des espaces, les liens sémantiques qui peuvent relier des mots entre eux, comme par exemple les mots "comédie" et "musicale", sont perdus.

Fonctionnement des méthodes d'analyse à thesaurus :

Ces méthodes sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus, cette définition étant, comme on l'a mentionné plus haut, entièrement manuelle et devant être opérée dans chaque domaine de spécialité.

Imaginons par exemple le thesaurus suivant :

spectacle → comédie (s) → dramatique
→ musicale → Cats
→ Les dix commandements
→ savante

Avec ce type de méthodes, il est toujours possible d'identifier les mots du texte source qui se retrouvent exactement sous la même forme dans le thesaurus. L'avantage de ces méthodes est que l'on peut être sûr que les mots identifiés correspondent à une réalité culturelle ou scientifique établie et répertoriée. D'autre part, il est possible de déduire un mot fédérateur comme "spectacle" qui ne fait pas partie du texte initial, mais qui le caractérise correctement. En revanche, l'inconvénient majeur de ces méthodes est qu'il faut perpétuellement mettre à jour le thesaurus pour qu'il conserve sa pertinence, ce qui entraîne des frais de maintenance importants. Un autre inconvénient important de ces méthodes réside dans le fait qu'un thesaurus constitué pour analyser des textes dans le domaine de la chimie ne pourra pas être utilisé pour des textes dans le domaine de l'électronique, par exemple. De plus, dans le cas où le thesaurus n'est pas exhaustif, certaines expressions qui

peuvent être très pertinentes ne seront pas reconnues comme telles.

Fonctionnement des méthodes d'analyse à reconnaissance de motifs :

Les méthodes d'analyse à reconnaissance de motifs que l'on connaît sont des méthodes d'identification statistiques de motifs qui, bien qu'elles améliorent considérablement les méthodes d'analyse statistique mentionnées plus haut, en conservant la trace de l'appariement des mots, comme par exemple des termes "comédie" et "musicale" de l'exemple ci-dessus, ne permettent pas d'analyser de façon correcte des textes courts. En effet, les méthodes statistiques ont besoin de quantité pour fonctionner correctement.

Par exemple, les motifs-clés du texte d'exemple seront obtenus par comparaisons approximatives de séquences plus ou moins longues entre elles. Les mots outils (le, la, les, …) ne comptent pas, et les séquences sont formées à partir d'un mot, plus ou moins trois mots :

Cats
Cats comédies
Cats comédies musicales
Cats comédies musicales longtemps
comédies
comédies musicales
comédies musicales longtemps
comédies musicales longtemps affiche
musicales
musicales longtemps
musicales longtemps affiche
musicales longtemps affiche tirer
etc…

Il suffit ensuite de regrouper les différentes séquences obtenues, par approximation sur la forme (par exemple

« comédies » et « comédie »), et de compter les expressions combinées les plus fréquentes comme « comédies musicales ».

Le but de la présente invention est de proposer un système pour l'extraction d'informations dans un texte en langage naturel permettant de remédier aux inconvénients des méthodes d'analyses connues, en permettant notamment une analyse de bonne qualité de textes aussi bien courts que longs.

Ce système utilise une méthode d'analyse par identification de motifs (patterns) non pas purement statistique, mais également syntaxique.

En résumé, le système proposé convertit les mots du texte en suite de catégories syntaxiques, puis confronte des sous-ensembles du texte avec des motifs syntaxiques prédéfinis, de façon à identifier des groupes nominaux sans préjuger de la valeur des mots qui composent ces groupes.

Ainsi, les mots « pomme de terre » ou « électronique de puissance » ne sont pas important par eux-mêmes, mais sont importants par rapport au texte où ils apparaissent. Dans un texte de nature générale « électronique de puissance » peut n'être qu'un exemple, pas un mot-clé du texte, mais sera probablement mot-clé dans un texte traitant des transistors. C'est le contexte qui fait le mot-clé, et le système selon la présente invention comporte en quelque sorte un analyseur de contextes syntaxiques. De même, le mot "porte" peut être reconnu comme nominal dans certains textes à cause de sa position par rapport aux autres mots du texte, ou simplement comme mot structurel dans d'autres textes.

Une méthode d'analyse par identification de motifs est proposée dans le document US 4,864,501. Le procédé décrit dans ce document antérieur utilise, pour le codage des mots du texte en vue de l'identification des motifs, un dictionnaire contenant les mots radicaux (base forms). Outre le fait que ce dictionnaire est très volumineux puisqu'il contient plusieurs

dizaines de milliers d'entrées, le procédé nécessite des algorithmes complexes de radicalisation des mots, spécifiques à chaque langue, pour retomber sur des mots du dictionnaire, ainsi qu'éventuellement des tables spécifiques de préfixes/suffixes pour traiter les cas d'erreurs d'orthographe, etc. Il s'agit par conséquent d'un procédé très lourd à mettre en œuvre et à utiliser.

Le système d'extraction selon la présente invention permet de remédier à ces inconvénients.

A cet effet, l'invention concerne un procédé d'extraction d'informations dans un texte en langage naturel, par identification de motifs (patterns), selon lequel on effectue un codage des mots du texte en les comparant avec le contenu d'un lexique prédéfini contenant quelques dizaines de mots outils, et selon lequel on identifie ensuite des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

L'invention concerne également un système d'extraction d'informations dans un texte en langage naturel comprenant :

- une unité d'entrée pour recevoir ledit texte en langage naturel,
- un fichier lexique dans lequel sont enregistrés des mots outils,
- un processeur d'analyse relié à ladite unité d'entrée, au fichier lexique et agencé pour effectuer dans un premier temps le codage des mots dudit texte en langage naturel par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit fichier lexique de mots outils, de façon d'une part à repérer les mots outils dans le texte et à évaluer la fonction des mots d'usage, non reconnus comme mots outils, en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils, et dans un deuxième temps une recherche, parmi

des sous-ensembles de la suite de mots codés obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies, de façon à identifier des groupes nominaux,

- une unité de sortie reliée audit processeur d'analyse pour recevoir les groupes de mots codés reconnus comme des motifs syntaxiques.

Le système d'extraction selon l'invention évalue la fonction grammaticale des mots du texte à analyser à l'aide d'un lexique prédéfini contenant les quelques dizaines de mots outils propres à chaque langue et qui sont essentiellement les articles, les prépositions, les conjonctions et auxiliaires verbaux. La fonction des autres mots est ensuite déduite grâce à l'emplacement des seuls mots outils. Du fait que les mots outils d'un texte représentent couramment 40 à 50 % des mots de ce texte, ceux-ci sont donc toujours assez nombreux pour permettre l'évaluation des autres mots. Ensuite, seules les parties du texte dont la grammaire est identifiée comme mots-clés possibles sont retenues.

Les avantages du système d'extraction selon l'invention sont nombreux. En particulier, le lexique de mots outils utilisé par le système est incomparablement plus léger que les dictionnaires contenant plusieurs milliers de mots qu'utilisent les systèmes connus. On relèvera, d'autre part, qu'aucune intervention humaine n'est nécessaire pour la détermination des mots-clés, que le système peut fonctionner pour des textes de langues diverses et que, mis à part le lexique des mots outils, il ne nécessite aucun autre lexique. De plus, du fait que la valeur sémantique et grammaticale des mots outils est fixe et n'évolue pratiquement jamais sur plusieurs décennies, la maintenance du lexique est des plus réduites. En revanche, la valeur des autres mots, que l'on peut appeler les mots d'usage (verbes, noms, adjectifs), évolue sans cesse dans le temps, en fonction des usages, de l'évolution des métiers ou des sciences, ou simplement en fonction de l'actualité. Du fait que le système de la présente

invention ne présuppose rien sur la valeur des mots d'usage, il fonctionne de façon identique dans tous les domaines, littéraire, technique ou scientifique, alors que les systèmes qui utilisent les méthodes connues doivent toujours être enrichis avec des lexiques spécialisés, fabriqués bien souvent sur mesure. Enfin, ce système d'extraction permet d'adresser de nouvelles langues incomparablement plus rapidement que n'importe quel autre système proposé jusqu'ici.

D'autre part, contrairement aux systèmes utilisant des méthodes d'analyse statistique dans lesquelles la fréquence d'apparition des mots est un critère de sélection, ce qui suppose que le texte soit suffisamment long, le système selon l'invention n'accorde à la fréquence d'apparition des mots qu'une importance subalterne et fonctionne aussi bien pour des textes longs de plusieurs dizaines de pages que pour des textes courts de quelques lignes.

On va décrire ci-après, à titre d'exemple, un système d'extraction d'informations selon l'invention dans un texte en langage naturel, en se référant aux dessins, sur lesquels :

- la fig. 1 est un schéma-bloc du système d'extraction selon l'invention;
- la fig. 2 est un schéma-bloc des étapes d'un mode d'exécution du procédé selon l'invention.

L'utilisation d'un modèle syntaxique requiert de reconnaître la langue du texte analysé. C'est donc naturellement la première opération qu'effectue le système d'extraction selon l'invention. Cette reconnaissance de la langue peut être basée sur des critères purement statistiques de cooccurrence de lettres. La reconnaissance des langues, par exemple anglais, espagnol, français, portugais, allemand ou italien, permet d'orienter les analyses qui seront réalisées en aval.

L'étape suivante est une étape de profilage du texte qui permet d'identifier les lignes de texte (paragraphes) comportant une information linguistique, et d'opérer des regroupements de paragraphes. Cette opération est particulièrement utile pour les textes structurés (avec titres, sous-titres, etc.), car elle permet de regrouper des paragraphes de façon cohérente. Elle est inutile pour des textes courts.

L'étape suivante consiste en une opération de régularisation du texte au cours de laquelle il s'agit d'éliminer les amalgames de signes, comme par exemple séparer les caractères typographiques des caractères alphabétiques. Il sera par exemple utile de reconnaître la chaîne "mot," comme le terme "mot" suivit de ",", alors que la chaîne "1,5" devra être reconnue comme un nombre.

Dans le texte d'exemple, cette étape revient à séparer les caractères typographiques (", " , "'" et ".") des autres mots par des espaces blancs. Le texte d'exemple devient alors :

"« Cats » , l ' une des comédies musicales les plus longtemps à l ' affiche , va tirer sa révérence après vingt et une années sur la scène londonienne . La dernière représentation de cette œuvre d ' Andrew Lloyd Webber aura lieu le 11 mai , jour de son 21e anniversaire , après quelque 9 000 représentations . L ' annonce a été faite trois jours après la dernière représentation de « Starlight Express » , la seconde comédie musicale la plus longtemps à l ' affiche à Londres , après dix-huit années sur les planches .
La fin de « Cats » est un coup dur supplémentaire pour le quartier de Covent Garden , où sont regroupés la plupart des théâtres londoniens , et qui a souffert d ' une forte baisse de fréquentation en 2001 . Depuis 1981 , année de son lancement , la comédie musicale a , depuis , été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays . "

L'étape suivante, qui constitue une étape clé du système, consiste à déterminer la catégorie de chaque mot. Grâce au lexique restreint des mots outils, les mots du texte sont codés selon des catégories grammaticales attribuées en fonction de la valeur syntaxique des mots. Les mots outils du lexique sont dans un premier temps reconnus dans le texte,

puis la fonction des autres mots du texte est déduite en fonction de leur emplacement par rapport aux mots outils déjà reconnus.

Ainsi, si l'on adopte par exemple les catégories suivantes :

s: mot de structure (mot outil non utile pour la suite de l'analyse)
d: déterminant (le, la, les, etc.)
p: préposition (de, en, par, etc.)
4: signe ouvrant ou fermant
1 ou 2 : ponctuation
3: apostrophe
N: nombre
W: nom propre
w: nom commun
c: amalgame (du, des, au, aux, …)
a: anaphores (ce, cet, ces, …)
*: code attribué si aucune des catégories précédentes n'est reconnue

Le texte d'exemple mentionné plus haut devient :

4 W 4 2 d 3 d c w3 w4 d w1 w2 p d 3 w3 2 s w2 a w4 w2 w1 p d w2 p d w2 w4 1 d w3 w5 p a w2 p 3 W W W w2 w1 d N w1 2 w1 p a * w5 2 w2 d N N w5 1 d 3 w3 s w2 w2 d w1 w2 d w3 w5 p 4 W W 4 2 d w3 w3 w4 d w1 w2 p d 3 w3 p W 2 w2 d 0 d w2 p d w2 1 d w1 p 4 W 4 s d w1 w1 w5 p d w2 p W W 2 s s w3 d w2 c w2 w3 2 p s s w2 p 3 d w2 w2 p w4 p N 1 W N 2 w2 p a w3 2 d w3 w4 s 2 w2 2 w2 w4 w2 w1 p N w2 p w3 p N w2 p p N w1 1

Une étape suivante consiste à identifier les structures linguistiques appelées syntagmes nominaux dans la terminologie linguistique ou, plus simplement, groupes nominaux.

L'ensemble des motifs syntaxiques qu'il est utile d'identifier constitue la grammaire d'analyse. Du fait que cette grammaire est commune à l'ensemble des langues romanes, il est possible d'analyser un grand nombre de langues en utilisant un même système d'extraction selon l'invention sans adaptation lourde.

A titre d'exemple, une grammaire (simplifiée) peut avoir la forme suivante :

(1) syntagme nominal -> déterminant , groupe nominal ; W .
(2) déterminant -> d ; d , 3 ; nombre ; c ; a
(3) d -> 'le' ; 'la' ; 'les' ; 'des' ; 'l' ; etc…
(3bis) c -> 'du' ; 'au' ; 'aux' ; etc…
(3ter) a -> 'ce' ; 'cette' ; 'ces' ; 'son' ; etc…
(4) groupe nominal -> expression , groupe nominal .
(5) expression -> w , p , w ; w .
(6) p -> 'de' ; 'à' ; 'pour' ; 'sans' ; etc…

La flèche se dit « se réécrit », la virgule se dit « suivi de », le point-virgule exprime un « ou », le point marque la fin de la règle. La règle (1) se lit « syntagme nominal se réécrit déterminant suivi de groupe nominal ».

Les règles (3) et (6) sont dites règles terminales car elles font appels aux formes lexicales du lexique des mots outils.

La règle (4) est une règle récursive. Un groupe nominal peut donc contenir une infinité d'expressions, lesquelles, selon la règle (5) sont soit de type wpw, soit de type w.

Les suites de catégories grammaticales suivantes seront donc reconnues comme syntagme nominal :

d w
d w p w
d w w
d w w p w
d 3 w w
etc…

Sur le texte d'exemple, les groupes nominaux identifiés à l'aide de cette grammaire ont été soulignés :

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelques 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.

(source Reuter)

Comme les groupes nominaux représentent à peu près 50 % du texte, il est nécessaire de ne retenir que ceux dont la probabilité d'être de vrais mots-clés du texte est la plus forte.

Une étape suivante peut consister à filtrer les groupes nominaux. Tous les groupes nominaux n'ont pas la même capacité référentielle. Certains sont plus importants que d'autres. Pour déterminer quels sont les plus importants d'entre eux, le système selon l'invention valorise chaque groupe nominal en fonction d'un double critère, l'un statistique, l'autre syntaxique.

Le critère statistique :

Les mots les plus fréquents des groupes nominaux sont classés par ordre de fréquence décroissant (en tenant compte d'une approximation comme 'comédie' = 'comédies'), soit dans le texte d'exemple :

comédie	3
musicale	3
affiche	2

années	2
Cats	2
dernière	2
représentation	2

Seuls les mots dont l'occurrence dépasse 1 sont conservés dans la liste. Les mots éliminés ont donc une valeur nulle. On ajoute à la valeur de chaque groupe nominal (initialement fixée à 0), la valeur de l'occurrence des mots qu'il contient moins 1. La valeur des groupes nominaux devient :

comédie musicale	(3 - 1) + (3 - 1) = 4
affiche	2 - 1 = 1
affiche à Londres	2 - 1 = 1
Cats	2 - 1 = 1
etc…	

Le critère syntaxique :

Lorsque qu'un groupe nominal est ou comporte un nom propre, celui-ci prend un point de valeur supplémentaire, 0 sinon.

comédie musicale	4 + 0 = 4
affiche	1 + 0 = 1
affiche à Londres	1 + 1 = 2
Cats	1 + 1 = 2
etc…	

Avec cette valorisation, il est aisé de procéder au classement des groupes nominaux. Dans le texte d'exemple, les groupes nominaux perçus comme les plus importants sont soulignés deux fois, les groupes d'importance secondaire sont soulignés une fois, tandis que les autres ont été purement et simplement éliminés.

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le

11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.
(source Reuter)

Revendications

1. Procédé d'extraction d'informations dans un texte en langage naturel, par identification de motifs (patterns), caractérisé en ce que l'on effectue un codage des mots du texte en les comparant avec le contenu d'un lexique prédéfini contenant quelques dizaines de mots outils, et en ce que l'on identifie ensuite des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

2. Procédé selon la revendication 1, caractérisé en ce que le codage des mots du texte s'effectue par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit lexique de mots outils, de façon à repérer les mots outils dans le texte et en ce que la fonction des mots d'usage, non reconnus comme mots outils, est déduite en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils.

3. Procédé selon l'une des revendications 1 ou 2, caractérisé en ce que les groupes nominaux identifiés sont ensuite valorisés de façon à ne retenir que les groupes perçus comme les plus importants en utilisant des critères de valorisation prédéfinis.

4. Système d'extraction d'informations dans un texte en langage naturel, caractérisé en ce qu'il comprend :

- une unité d'entrée pour recevoir ledit texte en langage naturel,
- un fichier lexique dans lequel sont enregistrés des mots outils,
- un processeur d'analyse relié à ladite unité d'entrée, au fichier lexique et agencé pour effectuer dans un premier temps le codage des mots dudit texte en langage naturel par évaluation de la fonction grammaticale de chaque mot en le comparant avec le

contenu dudit fichier lexique de mots outils, de façon d'une part à repérer les mots outils dans le texte et à évaluer la fonction des mots d'usage, non reconnus comme mots outils, en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils, et dans un deuxième temps une recherche, parmi des sous-ensembles de la suite de mots codés obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies, de façon à identifier des groupes nominaux,

- une unité de sortie reliée audit processeur d'analyse pour recevoir les groupes de mots codés reconnus comme des motifs syntaxiques.

5. Système selon la revendication 4, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de valorisation des groupes de mots codés retenus de façon à ne retenir que les groupes perçus comme les plus importants.

6. Système selon l'une des revendications 3 ou 4, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de reconnaissance de la langue du texte reçu dans l'unité d'entrée.

7. Système selon l'une des revendications 4 à 6, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de régularisation du texte reçu dans l'unité d'entrée de façon à éliminer les amalgames de signes.

Fig. 1

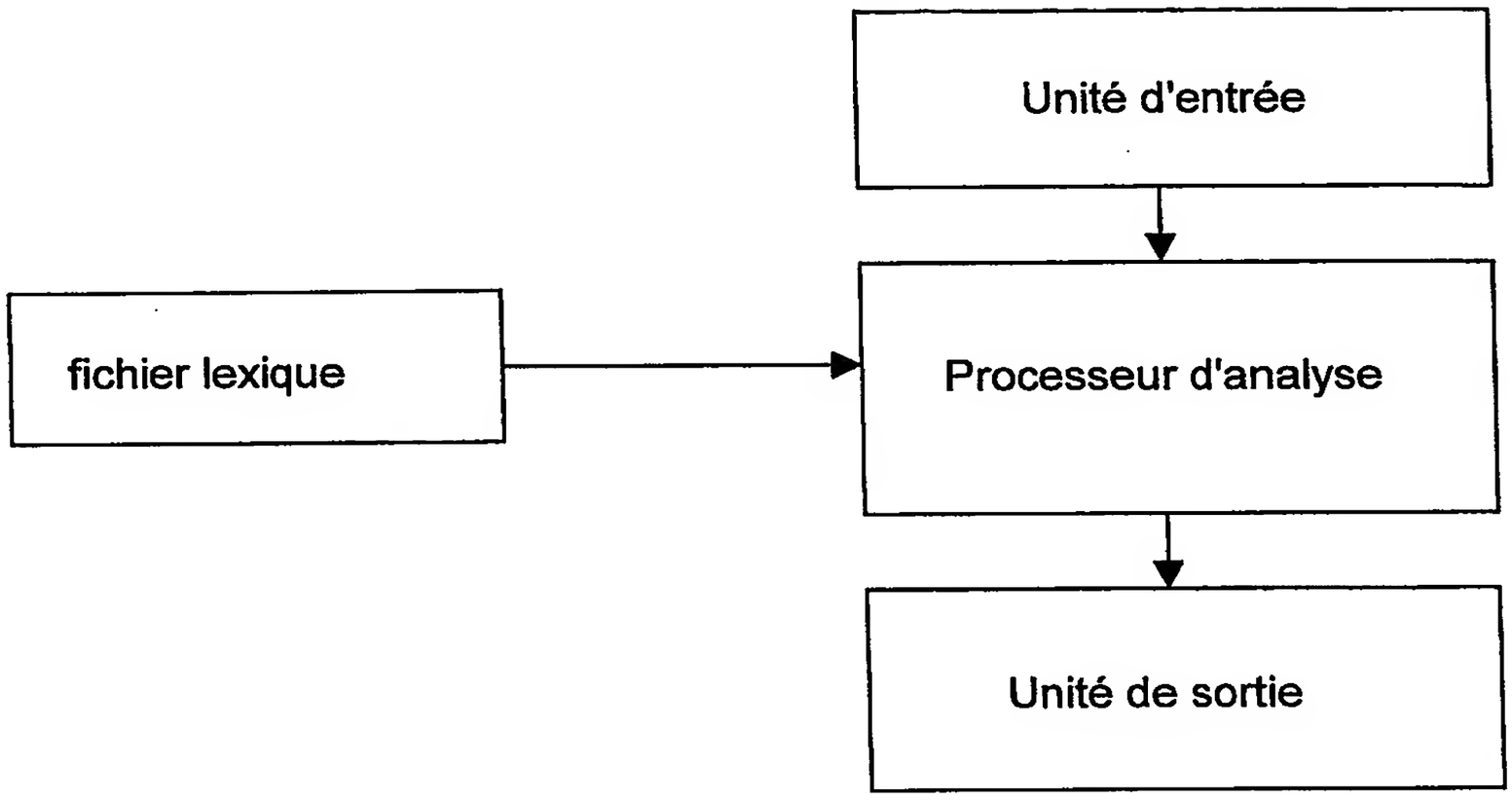
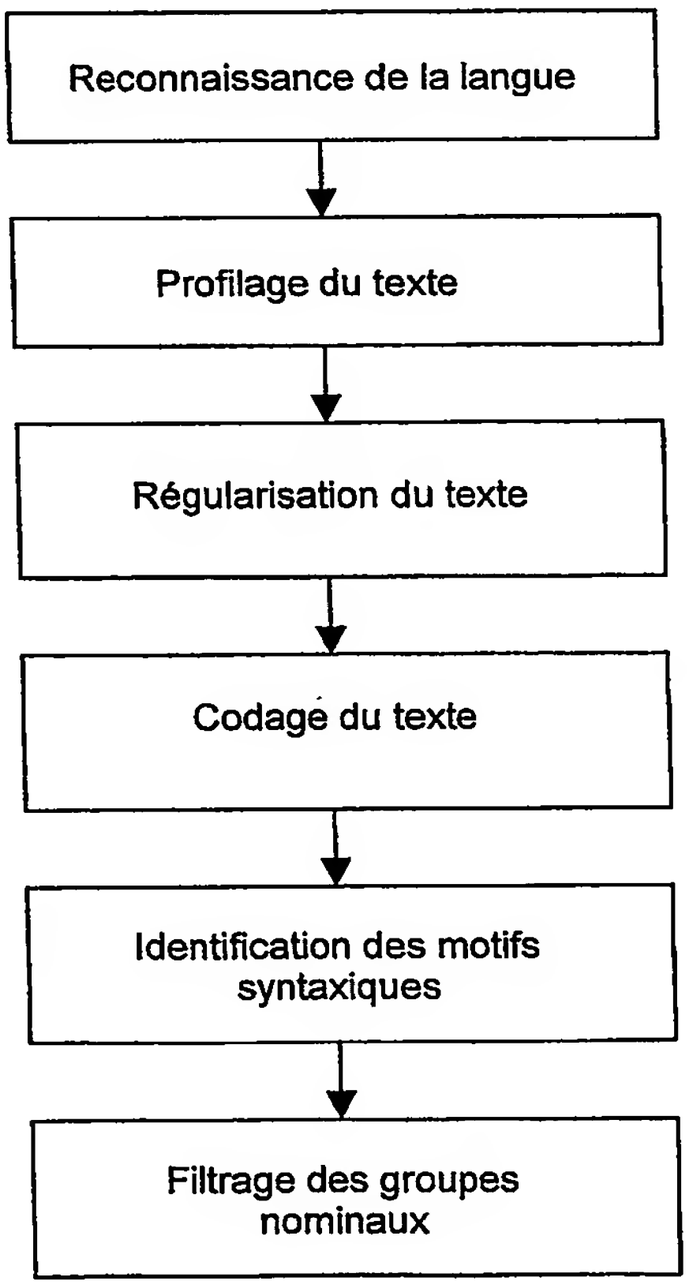

Unité d'entrée
fichier lexique
Processeur d'analyse
Unité de sortie

Fig. 2


Reconnaissance de la langue
Profilage du texte
Régularisation du texte
Codage du texte
Identification des motifs syntaxiques
Filtrage des groupes nominaux

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CH 03/00490

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/27

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 4 864 501 A (KUCERA HENRY ET AL) 5 September 1989 (1989-09-05) abstract column 3, paragraph 2 column 3, line 49 - line 57; figure 2 column 11, line 52 - line 64; figure 7	1 3
Y A	US 5 960 383 A (FLEISCHER ROBERT JOHN) 28 September 1999 (1999-09-28) abstract column 3, line 31 - line 45; figures 2,3	3 1,4,5
A	EP 0 702 289 A (OCE NEDERLAND BV) 20 March 1996 (1996-03-20) abstract	1,4,6

-/-

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 January 2004	26/01/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Barieux, M

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Application No
PCT/CH 03/00490

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 822 503 A (MATSUSHITA ELECTRIC IND CO LTD) 4 February 1998 (1998-02-04) abstract page 4, line 42 - line 45; figure 2	1,2,4

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/CH 03/00490

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
US 4864501 A	05-09-1989	CA	1300272 C	05-05-1992
US 5960383 A	28-09-1999	NONE		
EP 0702289 A	20-03-1996	FR	2723457 A1	09-02-1996
		CN	1125882 A	03-07-1996
		DE	69523848 D1	20-12-2001
		DE	69523848 T2	06-06-2002
		EP	0702289 A1	20-03-1996
		JP	3195522 B2	06-08-2001
		JP	8123636 A	17-05-1996
		US	5960113 A	28-09-1999
EP 0822503 A	04-02-1998	JP	3198932 B2	13-08-2001
		JP	10049543 A	20-02-1998
		EP	0822503 A1	04-02-1998

(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
29 janvier 2004 (29.01.2004)

PCT

(10) Numéro de publication internationale
WO 2004/010324 A3

(51) Classification internationale des brevets[7] : G06F 17/27

(21) Numéro de la demande internationale :
PCT/CH2003/000490

(22) Date de dépôt international : 18 juillet 2003 (18.07.2003)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
02405626.9 19 juillet 2002 (19.07.2002) EP

(71) Déposant *(pour tous les États désignés sauf US)* : **ALBERT-INC. S.A.** [CH/CH]; Rue Du Simplon 25, CH-1006 Lausanne (CH).

(72) Inventeur; et
(75) Inventeur/Déposant *(pour US seulement)* : **GERMAIN, Nicolas** [FR/FR]; 37, rue Jean-Pierre Bredy, F-69100 Villeurbanne (FR).

(74) Mandataires : **GANGUILLET, Cyril** etc.; Abrema Agence Brevets et Marques, Ganguillet & Humphrey, 16, Avenue du Théâtre, Case postale 2065, CH-1002 Lausanne (CH).

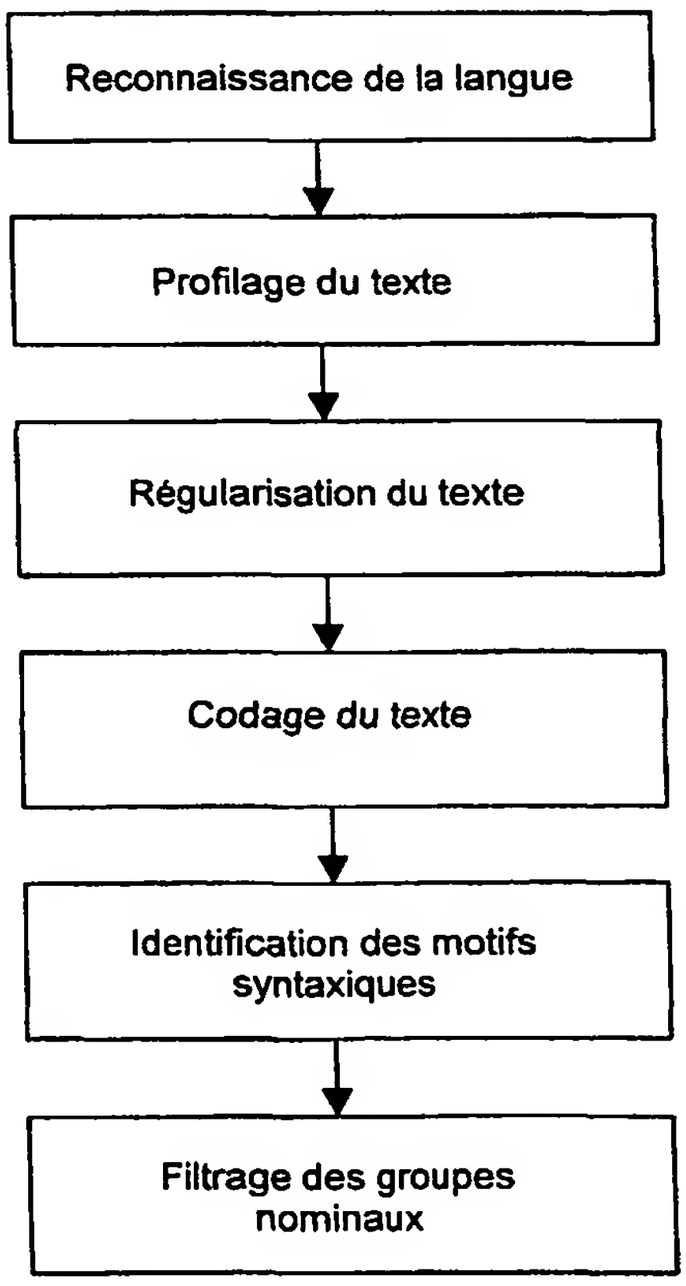
(81) États désignés *(national)* : AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) États désignés *(régional)* : brevet ARIPO (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), brevet

[Suite sur la page suivante]

(54) Title: SYSTEM FOR EXTRACTING INFORMATION FROM A NATURAL LANGUAGE TEXT

(54) Titre : SYSTEME D'EXTRACTION D'INFORMATIONS DANS UN TEXTE EN LANGAGE NATUREL



(57) Abstract: The invention relates to a system for extracting information from a natural language text. According to the invention, the extraction method consists in: encoding the words from the text by comparing said words with the contents of a lexicon of empty words (essentially articles, prepositions, conjunctions and verbal auxiliaries); and, subsequently, identifying noun phrases by searching for groups of encoded words that adhere to the pre-defined syntactic rules from among the subsets from the series of encoded words thus obtained.

(57) Abrégé : Le procédé d'extraction effectue un codage des mots du texte en les comparant avec le contenu d'un lexique de mots outils (essentiellement articles, prépositions, conjonctions et auxiliaires verbaux), puis identifie des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

WO 2004/010324 A3

eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), brevet européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), brevet OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

— *avec rapport de recherche internationale*

— *avant l'expiration du délai prévu pour la modification des revendications, sera republiée si des modifications sont reçues*

(88) Date de publication du rapport de recherche internationale: 1 avril 2004

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CH 03/00490

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/27

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 4 864 501 A (KUCERA HENRY ET AL) 5 September 1989 (1989-09-05) abstract column 3, paragraph 2 column 3, line 49 - line 57; figure 2 column 11, line 52 - line 64; figure 7 ---	1 3
Y A	US 5 960 383 A (FLEISCHER ROBERT JOHN) 28 September 1999 (1999-09-28) abstract column 3, line 31 - line 45; figures 2,3 ---	3 1,4,5
A	EP 0 702 289 A (OCE NEDERLAND BV) 20 March 1996 (1996-03-20) abstract --- -/--	1,4,6

[X] Further documents are listed in the continuation of box C. [X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 January 2004	26/01/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Barieux, M

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CH 03/00490

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 822 503 A (MATSUSHITA ELECTRIC IND CO LTD) 4 February 1998 (1998-02-04) abstract page 4, line 42 - line 45; figure 2	1,2,4

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/CH 03/00490

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 4864501 A	05-09-1989	CA 1300272 C	05-05-1992
US 5960383 A	28-09-1999	NONE	
EP 0702289 A	20-03-1996	FR 2723457 A1	09-02-1996
		CN 1125882 A	03-07-1996
		DE 69523848 D1	20-12-2001
		DE 69523848 T2	06-06-2002
		EP 0702289 A1	20-03-1996
		JP 3195522 B2	06-08-2001
		JP 8123636 A	17-05-1996
		US 5960113 A	28-09-1999
EP 0822503 A	04-02-1998	JP 3198932 B2	13-08-2001
		JP 10049543 A	20-02-1998
		EP 0822503 A1	04-02-1998

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale No
PCT/CH 03/00490

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
CIB 7 G06F17/27

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
CIB 7 G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si réalisable, termes de recherche utilisés)
EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie °	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X Y	US 4 864 501 A (KUCERA HENRY ET AL) 5 septembre 1989 (1989-09-05) abrégé colonne 3, alinéa 2 colonne 3, ligne 49 - ligne 57; figure 2 colonne 11, ligne 52 - ligne 64; figure 7	1 3
Y A	US 5 960 383 A (FLEISCHER ROBERT JOHN) 28 septembre 1999 (1999-09-28) abrégé colonne 3, ligne 31 - ligne 45; figures 2,3	3 1,4,5
A	EP 0 702 289 A (OCE NEDERLAND BV) 20 mars 1996 (1996-03-20) abrégé	1,4,6

-/--

[X] Voir la suite du cadre C pour la fin de la liste des documents

[X] Les documents de familles de brevets sont indiqués en annexe

° Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'inven tion revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'inven tion revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
19 janvier 2004	26/01/2004

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Barieux, M

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale No
PCT/CH 03/00490

C.(suite) DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie °	Identification des documents cités, avec,le cas échéant, l'indicationdes passages pertinents	no. des revendications visées
A	EP 0 822 503 A (MATSUSHITA ELECTRIC IND CO LTD) 4 février 1998 (1998-02-04) abrégé page 4, ligne 42 - ligne 45; figure 2	1,2,4

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale No
PCT/CH 03/00490

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 4864501 A	05-09-1989	CA 1300272 C	05-05-1992
US 5960383 A	28-09-1999	AUCUN	
EP 0702289 A	20-03-1996	FR 2723457 A1	09-02-1996
		CN 1125882 A	03-07-1996
		DE 69523848 D1	20-12-2001
		DE 69523848 T2	06-06-2002
		EP 0702289 A1	20-03-1996
		JP 3195522 B2	06-08-2001
		JP 8123636 A	17-05-1996
		US 5960113 A	28-09-1999
EP 0822503 A	04-02-1998	JP 3198932 B2	13-08-2001
		JP 10049543 A	20-02-1998
		EP 0822503 A1	04-02-1998

TRAITE DE COOPERATION EN MATIERE DE BREVETS

Expéditeur: le BUREAU INTERNATIONAL

PCT

NOTIFICATION DE L'ENREGISTREMENT D'UN CHANGEMENT

(règle 92bis.1 et instruction administrative 422 du PCT)

Destinataire:

GANGUILLET, Cyril
Abrema Agence Brevets et Marques
Ganguillet & Humphrey
16, Avenue du Théâtre
Case postale 2065
CH-1002 Lausanne
SUISSE

Date d'expédition (jour/mois/année) 17 août 2004 (17.08.2004)	
Référence du dossier du déposant ou du mandataire B-3851-WO	**NOTIFICATION IMPORTANTE**
Demande internationale no PCT/CH2003/000490	Date du dépôt international (jour/mois/année) 18 juillet 2003 (18.07.2003)

1. Les renseignements suivants étaient enregistrés en ce qui concerne:

[X] le déposant [ ] l'inventeur [ ] le mandataire [ ] le représentant commun

Nom et adresse	Nationalité (nom de l'Etat)	Domicile (nom de l'Etat)
ALBERT-INC. S.A. Rue Du Simplon 25 CH-1006 Lausanne SUISSE	CH	CH
	no de téléphone	
	no de télécopieur	
	no de téléimprimeur	

2. Le Bureau international notifie au déposant que le changement indiqué ci-après a été enregistré en ce qui concerne:

[X] la personne [ ] le nom [ ] l'adresse [ ] la nationalité [ ] le domicile

Nom et adresse	Nationalité (nom de l'Etat)	Domicile (nom de l'Etat)
GO ALBERT FRANCE SARL 12, rue Vivienne F-75002 Paris FRANCE	FR	FR
	no de téléphone	
	no de télécopieur	
	no de téléimprimeur	

3. Observations complémentaires, le cas échéant:

4. Une copie de cette notification a été envoyée:

[X] à l'office récepteur
[ ] à l'administration chargée de la recherche internationale
[ ] à l'administration chargée de l'examen préliminaire international
[X] aux offices désignés concernés
[ ] aux offices élus concernés
[ ] autre destinataire:

Bureau international de l'OMPI 34, chemin des Colombettes 1211 Genève 20, Suisse	Fonctionnaire autorisé: Marie-José DEVILLARD (Fax 338-8995)
no de télécopieur: (41-22) 338.89.95	no de téléphone: (41-22) 338 9439

Formulaire PCT/IB/306 (mars 1994)

006396799

CONTRAT

DE CESSION

DE DROITS DE PROPRIETE INTELLECTUELLE

entre

ALBERT-Inc. SA, société de droit suisse ayant son siège à la rue du Simplon 25, 1006 Lausanne (Suisse), représentée par M. Jacques Rosset, président, et Mme Beth Krasna, directrice générale

d'une part,

et

GO-ALBERT France, société à responsabilité limitée, au capital social de 100.000 €, dont le siège social est sis 12 Rue Vivienne, 75002 Paris, immatriculée au Registre du commerce et des sociétés de Paris sous le numéro RCS B 437 879 869, représentée par Madame Beth KRASNA et Monsieur Alain BEAUVIEUX, co-gérants,

ainsi que

M. Alain BEAUVIEUX, né le 13 juillet 1959 à Bagneux (France), domicilié 37 Avenue Daumesnil – 94160 SAINT MANDE, France

M. Eric FOURBOUL, né le 16 janvier 1967 à Villeneuve Saint Georges (France), domicilié 67 Rue Lunaret – 34000 MONTPELLIER, France

d'autre part.

* * *
* *
*

PREAMBULE

Albert-Inc. SA détient dans ses actifs différents brevets, marques et autres droits de propriété intellectuelle dans le domaine des nouvelles technologies.

Par convention du 2 et 3 septembre 2003, intitulée « Protocole de cession des parts sociales sous conditions suspensives », la société Albert-Inc. SA s'est engagée à céder la totalité de ses parts dans la société Go Albert France à M. Alain Beauvieux et à M. Eric Fourboul.

Cette convention est soumise à la condition suspensive de l'obtention d'un prêt maximum de 400.000 Euros d'une durée maximum de 4 mois et au taux maximum de 8 % par M. Alain Beauvieux et M. Eric Fourboul (article 8 du Protocole de cession des parts sociales sous conditions suspensives).

Afin de permettre à ces derniers de poursuivre l'activité d'Albert, Albert-Inc. SA consent à céder à Go Albert France les droits de propriété intellectuelle qu'elle détient.

La rémunération fixée d'entente entre les parties tient compte de la phase de liquidation dans laquelle Albert-Inc. SA est entrée et de l'objectif principal de l'opération qui vise à maintenir l'activité de la société Go Albert France.

CECI ETANT PRECISE, LES PARTIES CONVIENNENT CE QUI SUIT :

ARTICLE 1 : Objet de la vente

Pour les besoins du présent contrat, on entend par Droits de Propriété Intellectuelle l'ensemble des marques, dénominations, logos, noms de domaine, brevets, droits d'auteur, savoir-faire et autres droits similaires, notamment sur les programmes informatiques qu'ils fassent ou non l'objet de dépôts, demandes d'enregistrement ou enregistrements.

La société Albert-Inc. SA cède à la société Go Albert France tous les droits de propriété intellectuelle qu'elle détient, ainsi que les droits qui y sont attachés, notamment tous droits de priorité, sans aucune restriction de durée ni de territoire.

Il s'agit en particulier des marques et brevets dont la liste figure en Annexe 1.

ARTICLE 2 : Prix de vente

a) La société Go Albert France verse à la société Albert-Inc. SA le montant de un euro (EUR 1,-) au titre de contre-prestation de la cession des droits de propriété intellectuelle, ainsi que des droits qui y sont attachés.

b) Au cas où MM. Beauvieux et Fourboul et/ou Go Albert France vendent ou cèdent l'activité d'Albert et/ou la Propriété Intellectuelle acquise d'Albert-Inc. SA, sous quelque forme que ce soit, pour un montant supérieur à 10 millions d'euros dans un délai de 5 ans dès la signature de la présente convention, ces derniers s'engagent solidairement entre eux à verser un montant correspondant au 10 % du prix de vente ou de cession des activités d'Albert et/ou la Propriété Intellectuelle aux actionnaires et porteurs d'obligations d'Albert-Inc. SA à la date de l'Assemblée générale ordinaire du 2 septembre 2003.

et/ou GO ALBERT FRANCE

MM. Beauvieux et Fourboul s'engagent à tenir informé le représentant des actionnaires et porteurs d'obligations d'Albert-Inc. SA, en la personne de Me Jean-Philippe Rochat, à Lausanne, de toute vente ou cession d'activité ou de Propriété Intellectuelle intervenant dans les 5 ans, ainsi que des modalités auxquelles elle intervient. En outre ils s'engagent à lui adresser chaque année pendant ces 5 ans une copie signée des comptes annuels.

Au cas où les conditions posées ci-dessus sont réalisées, le montant dû aux actionnaires et porteurs de bons sera versé à Me Jean-Phillippe Rochat, à charge pour lui de verser ce montant à qui de droit.

pendant cette période de cinq ans

En tout temps, les actionnaires, par l'intermédiaire de leur représentant, peuvent obtenir de MM. Beauvieux et Fourboul et/ou de Go Albert France toute information relative à la vente ou à la cession de l'activité et/ou de la Propriété Intellectuelle.

ARTICLE 3 : Garanties

Au jour de la signature du présent contrat de cession, il n'existe, à la connaissance d'Albert-Inc. SA, pas de contestation des droits de propriété intellectuelle vendus. Aucune licence n'a été concédée pour l'utilisation de ces droits de propriété intellectuelle, à l'exception :

- des licences concédées par Albert-Inc. SA à ses filiales Go Albert France , Go Albert UK, Go Albert USA,

- des licences concédées par Albert-Inc. SA et les filiales précitées aux utilisateurs finaux.

Les droits de propriété intellectuelle sur les développements et créations réalisés par les filiales de Albert-Inc. SA ont été cédés à Albert-Inc. SA.

Albert-Inc. SA ne saurait être tenue pour responsable de contestations qui pourraient être soulevées par des tiers au sujet des droits de propriété intellectuelle vendus.

Albert-Inc. SA ne donne aucune garantie de la valeur des droits de propriété intellectuelle vendus.

Albert-Inc. SA ne donne aucune garantie relative au potentiel d'utilisation des droits de propriété intellectuelle vendus. Elle ne garantit en particulier pas que les droits de propriété intellectuelle vendus sont nécessaires et suffisants à l'exercice de l'activité commerciale actuellement déployée par Go Albert France.

ARTICLE 4 : Condition suspensive

Le présent contrat de cession est soumis à la condition suspensive de la réalisation de la cession de parts sociales prévues dans la convention principale liant Albert-Inc. SA et MM. Alain Beauvieux et Eric Fourboul.

ARTICLE 5 : Moment de la cession

La cession des droits de propriété intellectuelle vendus prend effet dès l'avènement de la condition suspensive visée à l'article 4 ci-dessus. Les droits issus des droits de propriété intellectuelle vendus passent alors de plein droit à Go Albert France.

Avant ce moment, Albert-Inc. SA conserve tous les droits attachés à ces droits de propriété intellectuelle. Elle s'engage toutefois à faire valoir ces droits de manière compatible avec le présent contrat et à ne rien entreprendre qui pourrait mettre son exécution en péril. Albert-Inc. SA s'engage en outre à prendre en charge tous les frais relatifs aux droits de propriété intellectuelle jusqu'à la date de cession effective.

ARTICLE 6 : Confidentialité et annonce

Les parties conviennent de se soumettre aux mêmes règles de confidentialité que celles régissant le Protocole de cession des parts sociales sous conditions suspensives (article 14 de ce Protocole).

Elle suivent les mêmes règles s'agissant de l'annonce de la cession des droits de propriété intellectuelle que celles contenues dans le Protocole de cession des parts sociales sous conditions suspensives (article 15 du Protocole).

ARTICLE 7 : Droit applicable et for

Le présent contrat est soumis au droit suisse.

Tout litige pouvant en découler sera soumis à la compétence exclusive des tribunaux du canton de Vaud.

ARTICLE 8 : Formalités

Albert-Inc. SA s'engage à fournir et signer tous documents nécessaires pour que Go Albert France puisse se prévaloir des droits de propriété intellectuelle acquis en vertu du présent contrat.

Les frais liés aux formalités d'inscription de la présente cession auprès des registres concernés seront à la charge de Go Albert France. Tous pouvoirs sont donnés au porteur d'un original des présentes aux fins d'accomplir lesdites formalités.

Ainsi fait en quatre exemplaires à Montpellier, le 15 septembre 2003

ALBERT-INC. SA

ALAIN BEAUVIEUX

GO ALBERT FRANCE

ERIC FOURBOUL

ANNEXE - LISTE DES MARQUES ET DES BREVETS DEVANT ETRE CEDES A LA SOCIETE ALBERT INC. PAR LA SOCIETE ALBERT-INC SA

liste des trademarks:

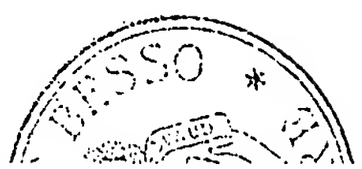
Exhaustive list of trademarks (incl. the trademark sign ™ or ® to use in corporate documentation)

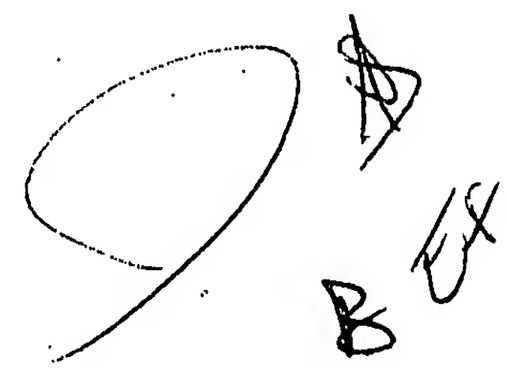
Albert™
AMI™
What You Mean Is What You Get™
meaning bus™

liste des brevets:

vous trouverez
ci-joint quatre documents (B-3561; B-3562; B-3563; B-3851) récapitulant l'état
des demandes de brevets concernant les quatre cas en cours :

- Enhancing Online Support;
- E-commerce using NLI;
- Natural Language Interface;
- GMIL.





Marque	Pays	Classe(s)	Date dépôt	No demande	Date enreg.	No enreg.	Ech. renouv.
ALBERT	Canada	= 9, 38, 42	20-févr-02	1131656			
	Etats-Unis	9	21-sept-00	78027141			
	Etats-Unis	42	5-oct-00	78029179			
	Marque communautaire	9, 38, 42	22-févr-02	2590354	17-juin-03	2590354	22-févr-12
	Marque internationale *	9, 38, 42	18-févr-02		1-mai-02	779 859	1-mai-12
	* Allemagne, Autriche, Benelux, Chine, Danemark, Espagne, Finlande, France, Italie, Japon, Norvège, Royaume-Uni, Suède						
	Suisse	9, 38, 42	18-févr-02	01495/2002	1-mai-02	498648	18-févr-12
ALBERT & FACE LOGO MARK	France	9, 38	28-sept-99	99814370	28-sept-99	99814370	28-sept-09
ALBERT MEANING INTERPRETER	Etats-Unis	9, 42	2-oct-01	78086540			
ALBERTS.COM	Etats-Unis	35, 42	31-janv-97	75234672	22-sept-98	2190417	22-sept-08
AMI	Canada	= 9, 38, 42	20-févr-02	1131657			
	Etats-Unis	9, 42	2-oct-01	78086549			
	Marque internationale *	9, 38, 42	18-févr-02		1-mai-02	779 861	1-mai-12
	* Allemagne, Benelux, France, Royaume-Uni						
	Suisse	9, 38, 42	18-févr-02	01498/2002	1-mai-02	498650	18-févr-12
FACE LOGO MARK	Canada	= 9, 38, 42	22-févr-02	1131998			

BESSO * GENIA

Marque	Pays	Classe(s)	Date dépôt	No demande	Date enreg.	No enreg.	Ech. renouv.
FACE LOGO MARK	Etats-Unis	9, 42	27-juin-00	78/014,566	17-sept-02	2,622,273	17-sept-12
	Marque communautaire	9, 38, 42	22-févr-02	2590065			
	Marque internationale *	9, 38, 42	18-févr-02		1-mai-02	779 862	1-mai-12
	* Allemagne, Autriche, Benelux, Chine, Danemark, Espagne, Finlande, France, Italie, Japon, Norvège, Royaume-Uni, Suède						
	Suisse	9, 38, 42	18-févr-02	01496/2002	1-mai-02	498653	18-févr-12
LIVO & FACE LOGO MARK	France	9, 35, 38, 42	5-juil-00	3038847	5-juil-00	3038847	5-juil-10
MEANING BUS	Canada	9, 42	28-mars-03	1172781			
	Etats-Unis	9, 42	25-mars-03	78229564			
	Marque internationale *	9, 42	2-oct-02		18-mars-03	800 787	18-mars-13
	* Allemagne, Benelux, France, Royaume-Uni						
	Suisse	9, 42	2-oct-02	08532/2002	18-mars-03	508197	2-oct-12
WHAT YOU MEAN IS WHAT YOU GET	Canada	= 9, 38, 42	20-févr-02	1131655			
	Etats-Unis	9, 42	3-août-01	78077331	21-janv-03	2,678,072	21-janv-13
	Marque internationale *	9, 38, 42	18-févr-02		1-mai-02	779 860	1-mai-12
	* Allemagne, Benelux, France, Royaume-Uni						
	Suisse	9, 38, 42	18-févr-02	01497/2002	1-mai-02	498649	18-févr-12

ABREMA

N/réf. : B-3561 Mai 2003

Titulaire : Albert-Inc. S.A. Demande de brevet : "**Enhancing Online Support**"

Pays	Date de dépôt	No de la demande	Date de délivrance	No du brevet	durée maximale
Etats-Unis	8 juin 1999	09/327,603	15 juillet 2003	6,594,657	8 juin 2019
Europe(*)	14 avril 2000	00915307.3	*Procédure d'examen en cours, pas encore de réaction de l'Office européen des brevets*		14 avril 2020
Canada	14 avril 2000	2,376,669	*Requête d'examen à déposer avant le 14 avril 2005*		14 avril 2020

(*) pays désignés : Allemagne, Autriche, Belgique, Chypre, Danemark, Espagne, Finlande, France, Royaume-Uni, Grèce, Irlande, Italie, Luxembourg, Monaco, Pays-Bas, Portugal, Suède, Suisse et Liechtenstein

ABREMA

N/réf. : B-3562 Mai 2003

Titulaire : Albert-Inc. S.A. Demande de brevet : **"E-commerce using NLI"**

Pays	Date de dépôt	No de la demande	Date de délivrance	No du brevet	durée maximale
Etats-Unis	8 juin 1999	09/327,604	3 septembre 2002	6,446,064	8 juin 2019
Europe(*)	14 avril 2000	00914337.1	*Procédure d'examen en cours, pas encore de réaction de l'Office européen des brevets*		14 avril 2020
Canada	14 avril 2000	2,376,671	*Requête d'examen à déposer avant le 14 avril 2005*		14 avril 2020

(*) pays désignés : Allemagne, Autriche, Belgique, Chypre, Danemark, Espagne, Finlande, France, Royaume-Uni, Grèce, Irlande, Italie, Luxembourg, Monaco, Pays-Bas, Portugal, Suède, Suisse et Liechtenstein

ABREMA

N/réf. : B-3563 Mai 2003

Titulaire : Albert-Inc. S.A. Demande de brevet : "**Natural Language Interface**"

Pays	Date de dépôt	No de la demande	Date de délivrance	No du brevet	durée maximale
Etats-Unis	8 juin 1999	09/327,605	22 juillet 2003	6,598,039	8 juin 2019
Europe(*)	14 avril 2000	00914338.9	*Procédure d'examen en cours, pas encore de réaction de l'Office européen des brevets*		14 avril 2020
Canada	14 avril 2000	2,376,672	*Requête d'examen à déposer avant le 14 avril 2005*		14 avril 2020

(*) pays désignés : Allemagne, Autriche, Belgique, Chypre, Danemark, Espagne, Finlande, France, Royaume-Uni, Grèce, Irlande, Italie, Luxembourg, Monaco, Pays-Bas, Portugal, Suède, Suisse et Liechtenstein

ABREMA

N/réf. : B-3851 Mai 2003

Titulaire : Albert-Inc. S.A. Demande de brevet : "**GMIL**"

Pays	Date de dépôt	No de la demande	Date de délivrance	No du brevet	durée maximale
Europe(*)	19 juillet 2002	02405626.9	*Requête d'examen à déposer dans les six mois qui suivront la publication du rapport de recherche, publication qui devrait intervenir avec la publication de la demande en janvier 2004.*		19 juillet 2022

(*) pays désignés : Allemagne, Autriche, Belgique, Chypre, Danemark, Espagne, Finlande, France, Royaume-Uni, Grèce, Irlande, Italie, Luxembourg, Monaco, Pays-Bas, Portugal, Suède, Suisse et Liechtenstein

Brevet No 1'546.- ——————————————————————————

Je soussigné Laurent Besso, notaire à Lausanne (Vaud - Suisse), pour le district de ce nom, atteste que la présente photocopie est conforme au document original qui m'a été présenté. ——————————————————————————

Lausanne, le vingt-neuf juillet deux mille cinq. ——————————————————————————

LAURENT BESSO * NOTAIRE *

LAURENT BESSO

CANTON DE VAUD LIBERTÉ ET PATRIE fr:2._

CANTON DE VAUD LIBERTÉ ET PATRIE fr:20._

CONTRACT

FOR THE ASSIGNMENT

OF INTELLECTUAL PROPERTY RIGHTS

between

ALBERT-Inc. SA, company under Swiss law having its head office at rue du Simplon 25, 1006 Lausanne (Switzerland), represented by Mr Jacques Rosset, Chairman, and Mrs Beth Krasna, General Manager

on the one hand

and

GO-ALBERT France, limited liability company, with share capital of €100,000, whose head office is situated at 12 Rue Vivienne, 75002 Paris, registered in the Paris Trade and Companies Register under the number RCS B 437 879 869, represented by Mrs Beth KRASNA and Mr Alain BEAUVIEUX, co-managers,

along with

Mr Alain BEAUVIEUX, born 13 July 1959 at Bagneux (France), residing at 37 Avenue Daumesnil, 94160 SAINT MANDE, France

Mr Eric FOURBOUL, born 16 January 1967 at Villeneuve Saint Georges (France), residing at 67 Rue Lunaret, 34000 MONTPELLIER, France

on the other hand.

***
**
*

PREAMBLE

Albert-Inc. SA holds amongst its assets various patents, trademarks and other intellectual property rights in the field of new technologies.

By an agreement of 2 and 3 September 2003, entitled "Heads of agreement for transfer of company shares under suspensive conditions", the company Albert-Inc. SA undertook to transfer all its shares in the company Go Albert France to Mr Alain Beauvieux and Mr Eric Fourboul.

This agreement is subject to the suspensive condition of the obtaining of a maximum loan of 400,000 euros for a maximum period of four months and at the maximum rate of 8% by Mr Alain Beauvieux and Mr Eric Fourboul (article 8 of the Heads of agreement on transfer of company shares under suspensive conditions).

In order to enable the latter to continue the activity of Albert, Albert-Inc. SA agrees to assign to Go Albert France the intellectual property rights that it holds.

The remuneration fixed by agreement between the parties takes account of the liquidation phase into which Albert-Inc. SA has entered and the main objective of the transaction, which aims to maintain the activity of the company Go Albert France.

THIS BEING STATED, THE PARTIES AGREE AS FOLLOWS:

ARTICLE 1: Object of the sale

For the requirements of the present contract, Intellectual Property Rights shall mean all the trademarks, names, logos, domain names, patents, copyrights, know-how and other similar rights, in particular over the computer programs, whether or not they are the subject of filings, applications for registration or registrations.

The company Albert-Inc. SA transfers to the company Go Albert France all the intellectual property rights that it holds as well as the rights attached thereto, in particular all priority rights, without any restriction on duration or territory.

It is a case in particular of the trademarks and patents, a list of which appears in Annex 1.

ARTICLE 2: Selling price

a) The company Go Albert France shall pay to the company Albert-Inc. SA the sum of one euro (€1.00) by way of consideration for the assignment of the intellectual property rights, and the rights attached thereto.

b) Should Messrs Beauvieux and Fourboul and/or Go Albert France sell or transfer the activity of Albert and/or the Intellectual Property acquired from Albert-Inc. SA, in any form whatsoever, for a sum greater than ten million euros within a period of five years from the signature of the present agreement, they undertake jointly to pay a sum corresponding to 10% of the selling or transfer price of the activities of Albert and/or the Intellectual Property to the shareholders and bondholders of Albert-Inc. SA at the date of the ordinary general meeting of 2 September 2003.

Messrs Beauvieux and Fourboul and/or Go Albert France undertake to keep the representative of the shareholders and bondholders of Albert-Inc. SA in the person of Maître Jean-Philippe Rochat, at Lausanne, informed of any sale or transfer of activity or Intellectual Property occurring within five years, and of the terms on which this occurs. In addition they undertake to send him each year during these five years a signed copy of the annual accounts.

Should the conditions posed above be fulfilled, the sum due to the shareholders and bondholders shall be paid to Maître Jean-Philippe Rochat, and it shall be incumbent on him to pay this amount to whoever is entitled.

At any time during this period of five years, the shareholders, through their representative, shall be entitled to obtain from Messrs Beauvieux and Fourboul and/or Go Albert France any information relating to the sale or transfer of the activity and/or of the Intellectual Property.

ARTICLE 3: Guarantees

At the date of signature of the present assignment contract, there does not exist, to the knowledge of Albert-Inc. SA, any dispute relating to the intellectual property rights sold. No licence has been granted for the use of these intellectual property rights, with the exception of:

- licences granted by Albert-Inc. SA to its subsidiaries Go Albert France, Go Albert UK, Go Albert USA,

- licences granted by Albert-Inc. SA and the aforementioned subsidiaries to the end users.

The intellectual property rights over the developments and creations made by the subsidiaries of Albert-Inc. SA have been assigned to Albert-Inc. SA.

Albert-Inc. SA shall not be held responsible for any disputes which may be raised by third parties with regard to the intellectual property rights sold.

Albert-Inc. SA gives no guarantee on the value of the intellectual property rights sold.

Albert-Inc. SA gives no guarantee relating to the potential for use of the intellectual property rights sold. In particular it does not guarantee that the intellectual property rights sold are necessary and sufficient for carrying on the commercial activity currently deployed by Go Albert France.

ARTICLE 4: Suspensive condition

The present assignment contract is subject to the suspensive condition of the carrying out of the transfer of company shares provided for in the main agreement between Albert-Inc. SA and Messrs Alain Beauvieux and Eric Fourboul.

ARTICLE 5: Time of transfer

The transfer of the intellectual property rights sold shall take effect along with the advent of the suspensive condition referred to in article 4 above. The rights issuing from the intellectual property rights sold shall then automatically pass to Go Albert France.

Before this time Albert-Inc. SA shall keep all the rights attached to these intellectual property rights. It undertakes however to assert these rights in a manner compatible with the present contract and to undertake nothing that might put its execution in danger. Albert-Inc. SA also undertakes to take responsibility for all the costs relating to the intellectual property rights until the effective date of transfer.

ARTICLE 6: Confidentiality and announcement

The parties agree to submit themselves to the same confidentiality rules as those governing the Heads of agreement on the transfer of company shares under suspensive conditions (article 14 of these Heads of agreement).

They shall follow the same rules with regard to the transfer of intellectual property rights as those contained in the Heads of agreement on the transfer of company shares under suspensive conditions (article 15 of the Heads of agreement).

ARTICLE 7: Applicable law and place of jurisdiction

The present contract shall be subject to Swiss law.

Any dispute that may stem from it shall come under the exclusive competence of the courts of the Canton of Vaud.

ARTICLE 8: Formalities

Albert-Inc. SA undertakes to provide and sign all documents necessary for Go Albert France to be able to prevail itself of the intellectual property rights acquired by virtue of the present contract.

The costs related to the formalities of registering the present transfer with the concerned registers shall be the responsibility of Go Albert France. All powers are given to the bearer of an original of these presents for the purpose of performing the said formalities.

Thus done in four copies in Montpellier, 15 September 2003.

[Signatures]

Albert-Inc. SA	Go Albert France
Alain Beauvieux	Eric Fourboul

ANNEX - LIST OF TRADEMARKS AND PATENTS TO BE ASSIGNED TO THE COMPANY ALBERT INC. BY THE COMPANY ALBERT-INC SA

List of trademarks:

Exhaustive list of trademarks (incl. the trademark sign ™ or ® to use in corporate documentation)

™

Albert™
AMI™
What You Mean Is What You Get™
meaning bus™

List of patents:

You will find enclosed herewith four documents (B-3561; B-3562; B-3563; B-3851) recapitulating the status of the patent applications concerning the four outstanding cases:

- Enhancing Online Support;
- E-commerce using NLI;
- Natural Language Interface;
- GMIL.

Mark	Country	Class(es)	Filing date	Filing No.	Registration date	Registration No.	Renewal term
ALBERT	Canada	= 9, 38, 42	Feb. 20, 2002	1131656			
	United States	9	Sept. 21, 2000	78027141			
	United States	42	Oct. 5, 2000	78029179			
	Community trademark	9, 38, 42	Feb. 22, 2002	2590354	June 17, 2003	2590354	Feb. 22, 2012
	International trademark*	9, 38, 42	Feb. 18, 2002		May 1, 2002	779 859	May 1, 2012
	* Germany, Austria, Benelux, China, Denmark, Spain, Finland, France, Italy, Japan, Norway, United Kingdom, Sweden						
	Switzerland	9, 38, 42	Feb. 18, 2002	01495/2002	May 1, 2002	498648	Feb. 18, 2012
ALBERT & FACE LOGO MARK	France	9, 38	Sept. 28, 1999	99814370	Sept. 28, 1999	99814370	Sept. 28, 2009
ALBERT MEANING INTERPRETER	United States	9, 42	Oct. 2, 2001	78086540			
ALBERTS.COM	United States	35, 42	Jan. 31, 1997	75234672	Sept. 22, 1998	2190417	Sept. 22, 2008
AMI	Canada	= 9, 38, 42	Feb. 20, 2002	1131657			
	United States	9, 42	Oct. 2, 2001	78086549			
	International trademark*	9, 38, 42	Feb. 18, 2002		May 1, 2002	779 861	May 1, 2012
	* Germany, Benelux, France, United Kingdom						
	Switzerland	9, 38, 42	Feb. 18, 2002	01498/2002	May 1, 2002	498650	Feb. 18, 2012

Marque	Pays	Classe(s)	Date dépôt	No demande	Date enreg.	No enreg.	Ech. renouv.
FACE LOGO MARK	Canada	= 9, 38, 42	Feb. 22, 2002	1131998			
	United States	9, 42	June 27, 2000	78/014,566	Sept. 17, 2002	2,622,273	Sept. 17, 2012
	Community trademark	9, 38, 42	Feb. 22, 2002	2590065			
	International trademark*	9, 38, 42	Feb. 18, 2002		May 1, 2002	779 862	May 1, 2012
	* Germany, Austria, Benelux, China, Denmark, Spain, Finland, France, Italy, Japan, Norway, United Kingdom, Sweden						
	Switzerland	9, 38, 42	Feb. 18, 2002	01496/2002	May 1, 2002	498653	Feb. 18, 2012
LIVO & FACE LOGO MARK	France	9, 35, 38, 42	July 5, 2000	3038847	July 5, 2000	3038847	July 5, 2010
MEANING BUS	Canada	9, 42	March 28, 2003	1172781			
	United States	9, 42	March 25, 2003	78229564			
	International trademark*	9, 42	Oct. 2, 2002		March 18, 2003	800 787	March 18, 2013
	* Germany, Benelux, France, United Kingdom						
	Switzerland	9, 42	Oct. 2, 2002	08532/2002	March 18, 2003	508197	Oct. 2, 2012
WHAT YOU MEAN IS WHAT YOU GET	Canada	= 9, 38, 42	Feb. 20, 2002	1131655			
	United States	9, 42	Aug. 3, 2001	78077331	Jan. 21, 2003	2,678,072	Jan. 21, 2013
	International trademark*	9, 38, 42	Feb. 18, 2002		May 1, 2002	779 860	May 1, 2012
	* Germany, Benelux, France, United Kingdom						
	Switzerland	9, 38, 42	Feb. 18, 2002	01497/2002	May 1, 2002	498649	Feb. 18, 2012

ABREMA

O/ref. : B-3561

May 2003

Owner : Albert-Inc. S.A.

Patent application: **"Enhancing Online Support"**

Country	Filing date	Filing No.	Date of grant	Patent No.	Maximum duration
United States	June 8, 1999	09/327,603	July 15, 2003	6,594,657	June 8, 2019
Europe(*)	April 14, 2000	00915307.3	*Examination pending, no reaction of the European Patent Office yet.*		April 14, 2020

(*) designated countries : Germany, Austria, Belgium, Cyprus, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Luxembourg, Monaco, the Netherlands, Portugal, Sweden, Switzerland and Liechtenstein

Canada	April 14, 2000	2,376,669	*Examination request to be filed before April 14, 2005.*		April 14, 2020

ABREMA

O/ref. : B-3562 May 2003

Owner : Albert-Inc. S.A. Patent application: **"E-commerce using NLI"**

Country	Filing date	Filing No.	Date of grant	Patent No.	Maximum duration
United States	June 8, 1999	09/327,604	September 3, 2002	6,446,064	June 8, 2019
Europe(*)	April 14, 2000	00914337.1	*Examination pending, no reaction of the European Patent Office yet.*		April 14, 2020

(*) designated countries : Germany, Austria, Belgium, Cyprus, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Luxembourg, Monaco, the Netherlands, Portugal, Sweden, Switzerland and Liechtenstein

Canada	April 14, 2000	2,376,671	*Examination request to be filed before April 14, 2005.*		April 14, 2020

ABREMA

O/ref. : B-3563

May 2003

Owner : Albert-Inc. S.A.

Patent application : **"Natural Language Interface"**

Country	Filing date	Filing No.	Date of grant	Patent No.	Maximum duration
United States	June 8, 1999	09/327,605	July 22, 2003	6,598,039	June 8, 2019
Europe(*)	April 14, 2000	00914338.9	*Examination pending, no reaction of the European Patent Office yet.*		April 14, 2020

(*) designated countries : Germany, Austria, Belgium, Cyprus, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Luxembourg, Monaco, the Netherlands, Portugal, Sweden, Switzerland and Liechtenstein

Canada	April 14, 2000	2,376,672	*Examination request to be filed before April 14, 2005.*		April 14, 2020

ABREMA

O/ref. : B-3851

May 2003

Owner : Albert-Inc. S.A.

Patent application : **"GMIL"**

Country	Filing date	Filing No.	Date of grant	Patent No.	Maximum duration
Europe(*)	July 19, 2002	02405626.9	*Examination request to be filed within six months following the publication of the search report, such publication being to occur with the publication of the application in January 2004.*		July 19, 2022

(*) designated countries : Germany, Austria, Belgium, Cyprus, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Luxembourg, Monaco, the Netherlands, Portugal, Sweden, Switzerland and Liechtenstein

English translation of the Certification of the Notary public:

Case No. 1545

I, the undersigned Laurent Besso, Notary public in Lausanne (Vaud – Switzerland), for the district of the same name, certifiy that the present copy is a true copy of the original document which has been presented to me.

Lausanne, July 29, 2005

Stamp of Laurent Besso, Notary public - Signature

ASSIGNMENT OF INVENTION

In consideration of the payment by ASSIGNEE to ASSIGNOR of the sum of One Dollar ($ 1.00), the receipt of which is hereby acknowledged, and for other good and valuable consideration,

ASSIGNOR: **Nicolas Germain**
37, rue Jean-Pierre Bredy
69100 Villeurbanne
France

hereby sells, assigns and transfers to

ASSIGNEE: **Albert-Inc. S.A.**
Rue du Simplon 25
1006 Lausanne
Switzerland

and the successors, assigns and legal representatives of the ASSIGNEE the entire right, title and interest for the whole world, including the United States of America and its territorial possessions, in and to any and all improvements which are disclosed, described and claimed or intended so to be in the invention entitled:

SYSTEME D'EXTRACTION D'INFORMATIONS DANS UN TEXTE EN LANGAGE NATUREL

which is described in the attached specification and in any and all patent applications to be filed thereon and arising therefrom, and in, to and under all Letters Patent to be obtained for said invention by the above application or by said other applications or by any application claiming priority thereof, or any continuation, division, renewal, or substitue thereof, and as to Letters Patent any re-issue or re-examination thereof.

ASSIGNOR hereby covenants that no assignment, sale, agreement or encumbrance has been or will be made or entered into which would conflict with this assignment.

ASSIGNOR further covenants that ASSIGNEE will, upon its request, be provided promptly with all pertinent facts and documents relating to said invention and said Letters Patent and legal equivalents as may be known and accessible to ASSIGNOR and will testify as to the same in any interference, litigation or proceeding related thereto and will promptly execute and deliver to ASSIGNEE or ASSIGNEE's legal representatives any and all papers, instruments or affidavits required to apply for, obtain, maintain, issue and enforce said application, said invention and said Letters Patent and said equivalents thereof which may be necessary or desirable to carry out the purposes thereof.

IN WITNESS WHEREOF, ASSIGNOR has hereunto set hand this 28 day of June, 2002.

Nicolas Germain

Legalization of the signature by a Notary Public:

JE SOUSSIGNÉ Maitre Dominique Grasset
Membre de la S.C.P. Jacques de BENOIST DE LA PRUNAREDE
et Dominique GRASSET, Notaires Associés titulaire d'un
Office notarial à la Résidence de BAILLARGUES (Hérault)
certifie que la signature apposée ci-dessus est bien celle de
M. Nicolas Germain
A BAILLARGUES
Le 28 juin 2002

Dominique GRASSET Notaire Associé
BAILLARGUES (Hérault)

Système d'extraction d'informations dans un texte en langage naturel

La présente invention concerne un système d'extraction d'informations dans un texte en langage naturel, en vue de sélectionner les mots ou les groupes de mots du texte qui décrivent le mieux les sujets abordés dans le texte. Ces mots ou groupes de mots sont appelés les "mots-clés" et sont notamment utilisables à des fins d'indexation du texte dans une base de données documentaire, en particulier pour le résumé automatique du texte, pour la catégorisation ou toute autre tentative de représentation de la connaissance.

Les systèmes d'extraction d'informations que l'on connaît et qui tentent d'atteindre ces objectifs utilisent des méthodes d'analyses de trois types :

- les méthodes d'analyse statistique qui tentent d'élire les mots du texte les plus représentatifs en comptant leurs fréquences d'apparition et en ne retenant que ceux dont la fréquence n'est ni trop faible, ni trop forte;

- les méthodes d'analyse à thesaurus qui fonctionnent d'après une représentation prédéfinie de la connaissance et qui sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus. Cette définition est entièrement manuelle et doit être opérée dans chaque domaine de spécialités;

- les méthodes d'analyse à reconnaissance de motifs (patterns) qui fonctionnent à l'aide d'identifications statistiques de motifs (patterns).

Le fonctionnement comparatif de ces trois types de méthodes d'analyse va être illustré ci-après par l'analyse du texte suivant :

"«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays."

(source Reuter)

Fonctionnement des méthodes d'analyse statistique :

Si l'on considère leur approche de façon caricaturale, les méthodes d'analyse statistique comptent les mots du texte pour ne retenir que ceux dont la fréquence n'est ni trop faible ni trop forte en éliminant parfois les mots outils (articles, prépositions, conjonctions, auxiliaires verbaux), afin d'affiner les résultats. En ce qui concerne le texte proposé ci-dessus, les mots "moyennement" fréquents (sans prendre en considération les mots outils) sont alors :

affiche, années, Cats, comédie, dernière, été, longtemps, musicale et représentation.

Bien que le principal avantage des méthodes d'analyse statistique réside dans une grande simplicité algorithmique, leur principal désavantage réside en la faible pertinence des résultats. En effet, les mots "moyennement" fréquents d'un texte sont rarement les plus représentatifs. Ces méthodes peuvent toutefois donner de meilleurs résultats sur des textes plus longs que le texte d'exemple ci-dessus.

D'autre part, du fait que le texte est découpé en mots, c'est-à-dire en chaînes de caractères dont les délimiteurs sont des espaces, les liens sémantiques qui peuvent relier des mots entre eux, comme par exemple les mots "comédie" et "musicale", sont perdus.

Fonctionnement des méthodes d'analyse à thesaurus :

Ces méthodes sont basées sur la définition préalable d'un lexique structuré de référence appelé thesaurus, cette définition étant, comme on l'a mentionné plus haut, entièrement manuelle et devant être opérée dans chaque domaine de spécialité.

Imaginons par exemple le thesaurus suivant :

spectacle → comédie (s) → dramatique
→ musicale → Cats
→ Les dix commandements
→ savante

Avec ce type de méthodes, il est toujours possible d'identifier les mots du texte source qui se retrouvent exactement sous la même forme dans le thesaurus. L'avantage de ces méthodes est que l'on peut être sûr que les mots identifiés correspondent à une réalité culturelle ou scientifique établie et répertoriée. D'autre part, il est possible de déduire un mot fédérateur comme "spectacle" qui ne fait pas partie du texte initial, mais qui le caractérise correctement. En revanche, l'inconvénient majeur de ces méthodes est qu'il faut perpétuellement mettre à jour le thesaurus pour qu'il conserve sa pertinence, ce qui entraîne des frais de maintenance importants. Un autre inconvénient important de ces méthodes réside dans le fait qu'un thesaurus constitué pour analyser des textes dans le domaine de la chimie ne pourra pas être utilisé pour des textes dans le domaine de l'électronique, par exemple. De plus, dans le cas où le thesaurus n'est pas exhaustif, certaines expressions qui

peuvent être très pertinentes ne seront pas reconnues comme telles.

Fonctionnement des méthodes d'analyse à reconnaissance de motifs :

Les méthodes d'analyse à reconnaissance de motifs que l'on connaît sont des méthodes d'identification statistiques de motifs qui, bien qu'elles améliorent considérablement les méthodes d'analyse statistique mentionnées plus haut, en conservant la trace de l'appariement des mots, comme par exemple des termes "comédie" et "musicale" de l'exemple ci-dessus, ne permettent pas d'analyser de façon correcte des textes courts. En effet, les méthodes statistiques ont besoin de quantité pour fonctionner correctement.

Par exemple, les motifs-clés du texte d'exemple seront obtenus par comparaisons approximatives de séquences plus ou moins longues entre elles. Les mots outils (le, la, les, …) ne comptent pas, et les séquences sont formées à partir d'un mot, plus ou moins trois mots :

Cats
Cats comédies
Cats comédies musicales
Cats comédies musicales longtemps
comédies
comédies musicales
comédies musicales longtemps
comédies musicales longtemps affiche
musicales
musicales longtemps
musicales longtemps affiche
musicales longtemps affiche tirer
etc…

Il suffit ensuite de regrouper les différentes séquences obtenues, par approximation sur la forme (par exemple

« comédies » et « comédie »), et de compter les expressions combinées les plus fréquentes comme « comédies musicales ».

Le but de la présente invention est de proposer un système pour l'extraction d'informations dans un texte en langage naturel permettant de remédier aux inconvénients des méthodes d'analyses connues, en permettant notamment une analyse de bonne qualité de textes aussi bien courts que longs.

A cet effet, le système utilise une méthode d'analyse par identification de motifs (patterns) non pas statistique, mais syntaxique.

En résumé, le système proposé convertit les mots du texte en suite de catégories syntaxiques, puis confronte des sous-ensembles du texte avec des motifs syntaxiques prédéfinis, de façon à identifier des groupes nominaux sans préjuger de la valeur des mots qui composent ces groupes.

Ainsi, les mots « pomme de terre » ou « électronique de puissance » ne sont pas important par eux-mêmes, mais sont importants par rapport au texte où ils apparaissent. Dans un texte de nature générale « électronique de puissance » peut n'être qu'un exemple, pas un mot-clé du texte, mais sera probablement mot-clé dans un texte traitant des transistors. C'est le contexte qui fait le mot-clé, et le système selon la présente invention comporte en quelque sorte un analyseur de contextes syntaxiques. De même, le mot "porte" peut être reconnu comme nominal dans certains textes à cause de sa position par rapport aux autres mots du texte, ou simplement comme mot structurel dans d'autres textes.

Le système d'extraction selon l'invention évalue la fonction grammaticale des mots du texte à analyser à l'aide d'un lexique prédéfini contenant les quelques dizaines de mots outils propres à chaque langue et qui sont essentiellement les articles, les prépositions, les conjonctions et auxiliaires verbaux. La fonction des autres mots est ensuite déduite grâce

à l'emplacement des seuls mots outils. Du fait que les mots outils d'un texte représentent couramment 40 à 50 % des mots de ce texte, ceux-ci sont donc toujours assez nombreux pour permettre l'évaluation des autres mots. Ensuite, seules les parties du texte dont la grammaire est identifiée comme mots-clés possibles sont retenues.

Les avantages du système d'extraction selon l'invention sont nombreux. On relèvera, en particulier, qu'aucune intervention humaine n'est nécessaire pour la détermination des mots-clés, que le système peut fonctionner pour des textes de langues diverses et que, mis à part le lexique des mots outils, il ne nécessite aucun autre lexique. De plus, du fait que la valeur sémantique et grammaticale des mots outils est fixe et n'évolue pratiquement jamais sur plusieurs décennies, la maintenance du lexique est des plus réduites. En revanche, la valeur des autres mots, que l'on peut appeler les mots d'usage (verbes, noms, adjectifs), évolue sans cesse dans le temps, en fonction des usages, de l'évolution des métiers ou des sciences, ou simplement en fonction de l'actualité. Du fait que le système de la présente invention ne présuppose rien sur la valeur des mots d'usage, il fonctionne de façon identique dans tous les domaines, littéraire, technique ou scientifique, alors que les systèmes qui utilisent les méthodes connues doivent toujours être enrichis avec des lexiques spécialisés, fabriqués bien souvent sur mesure.

D'autre part, contrairement aux systèmes utilisant des méthodes d'analyse statistique dans lesquelles la fréquence d'apparition des mots est un critère de sélection, ce qui suppose que le texte soit suffisamment long, le système selon l'invention n'accorde à la fréquence d'apparition des mots qu'une importance subalterne et fonctionne aussi bien pour des textes longs de plusieurs dizaines de pages que pour des textes courts de quelques lignes.

On va décrire ci-après, à titre d'exemple, un système d'extraction d'informations selon l'invention dans un texte en langage naturel, en se référant aux dessins, sur lesquels :

- la fig. 1 est un schéma-bloc du système d'extraction selon l'invention;

- la fig. 2 est un schéma-bloc des étapes d'un mode d'exécution du procédé selon l'invention.

L'utilisation d'un modèle syntaxique requiert de reconnaître la langue du texte analysé. C'est donc naturellement la première opération qu'effectue le système d'extraction selon l'invention. Cette reconnaissance de la langue peut être basée sur des critères purement statistiques de cooccurrence de lettres. La reconnaissance des langues, par exemple anglais, espagnol, français, portugais, allemand ou italien, permet d'orienter les analyses qui seront réalisées en aval.

L'étape suivante est une étape de profilage du texte qui permet d'identifier les lignes de texte (paragraphes) comportant une information linguistique, et d'opérer des regroupements de paragraphes. Cette opération est particulièrement utile pour les textes structurés (avec titres, sous-titres, etc.), car elle permet de regrouper des paragraphes de façon cohérente. Elle est inutile pour des textes courts.

L'étape suivante consiste en une opération de régularisation du texte au cours de laquelle il s'agit d'éliminer les amalgames de signes, comme par exemple séparer les caractères typographiques des caractères alphabétiques. Il sera par exemple utile de reconnaître la chaîne "mot," comme le terme "mot" suivit de ",", alors que la chaîne "1,5" devra être reconnue comme un nombre.

Dans le texte d'exemple, cette étape revient à séparer les caractères typographiques (", " , "'" et ".") des autres mots par des espaces blancs. Le texte d'exemple devient alors :

"« Cats » , l ' une des comédies musicales les plus longtemps à l ' affiche , va tirer sa révérence après vingt et une années sur la scène londonienne . La dernière représentation de cette œuvre d ' Andrew Lloyd Webber aura lieu le 11 mai , jour de son 21e anniversaire , après quelque 9 000 représentations . L ' annonce a été faite trois jours après la dernière représentation de « Starlight Express » , la seconde comédie musicale la plus longtemps à l ' affiche à Londres , après dix - huit années sur les planches .
La fin de « Cats » est un coup dur supplémentaire pour le quartier de Covent Garden , où sont regroupés la plupart des théâtres londoniens , et qui a souffert d ' une forte baisse de fréquentation en 2001 . Depuis 1981 , année de son lancement , la comédie musicale a , depuis , été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays . "

L'étape suivante, qui constitue une étape clé du système, consiste à déterminer la catégorie de chaque mot. Grâce au lexique restreint des mots outils, les mots du texte sont codés selon des catégories grammaticales attribuées en fonction de la valeur syntaxique des mots. Les mots outils du lexique sont dans un premier temps reconnus dans le texte, puis la fonction des autres mots du texte est déduite en fonction de leur emplacement par rapport aux mots outils déjà reconnus.

Ainsi, si l'on adopte par exemple les catégories suivantes :

s: mot de structure (mot outil non utile pour la suite de l'analyse)
d: déterminant (le, la, les, etc.)
p: préposition (de, en, par, etc.)
4: signe ouvrant ou fermant
1 ou 2 : ponctuation
3: apostrophe
N: nombre
W: nom propre
w: nom commun
c: amalgame (du, des, au, aux, ...)

a: anaphores (ce, cet, ces, ...)

*: code attribué si aucune des catégories précédentes n'est reconnue

Le texte d'exemple mentionné plus haut devient :

4 W 4 2 d 3 d c w3 w4 d w1 w2 p d 3 w3 2 s w2 a w4 w2 w1 p d w2 p d w2 w4 1 d w3 w5 p a w2 p 3 W W W w2 w1 d N w1 2 w1 p a * w5 2 w2 d N N w5 1 d 3 w3 s w2 w2 d w1 w2 d w3 w5 p 4 W W 4 2 d w3 w3 w4 d w1 w2 p d 3 w3 p W 2 w2 d 0 d w2 p d w2 1 d w1 p 4 W 4 s d w1 w1 w5 p d w2 p W W 2 s s w3 d w2 c w2 w3 2 p s s w2 p 3 d w2 w2 p w4 p N 1 W N 2 w2 p a w3 2 d w3 w4 s 2 w2 2 w2 w4 w2 w1 p N w2 p w3 p N w2 p p N w1 1

Une étape suivante consiste à identifier les structures linguistiques appelées syntagmes nominaux dans la terminologie linguistique ou, plus simplement, groupes nominaux.

L'ensemble des motifs syntaxiques qu'il est utile d'identifier constitue la grammaire d'analyse. Du fait que cette grammaire est commune à l'ensemble des langues romanes, il est possible d'analyser un grand nombre de langues en utilisant un même système d'extraction selon l'invention sans adaptation lourde.

A titre d'exemple, une grammaire (simplifiée) peut avoir la forme suivante :

(1) syntagme nominal -> déterminant , groupe nominal ; W .
(2) déterminant -> d ; d , 3 ; nombre ; c ; a
(3) d -> 'le' ; 'la' ; 'les' ; 'des' ; 'l' ; etc...
(3bis) c -> 'du' ; 'au' ; 'aux' ; etc...
(3ter) a -> 'ce' ; 'cette' ; 'ces' ; 'son' ; etc...
(4) groupe nominal -> expression , groupe nominal .
(5) expression -> w , p , w ; w .
(6) p -> 'de' ; 'à' ; 'pour' ; 'sans' ; etc...

La flèche se dit « se réécrit », la virgule se dit « suivi de », le point-virgule exprime un « ou », le point marque la fin de la règle. La règle (1) se lit « syntagme nominal se réécrit déterminant suivi de groupe nominal ».

Les règles (3) et (6) sont dites règles terminales car elles font appels aux formes lexicales du lexique des mots outils.

La règle (4) est une règle récursive. Un groupe nominal peut donc contenir une infinité d'expressions, lesquelles, selon la règle (5) sont soit de type wpw, soit de type w.

Les suites de catégories grammaticales suivantes seront donc reconnues comme syntagme nominal :

d w
d w p w
d w w
d w w p w
d 3 w w
etc…

Sur le texte d'exemple, les groupes nominaux identifiés à l'aide de cette grammaire ont été soulignés :

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelques 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.

(source Reuter)

Comme les groupes nominaux représentent à peu près 50 % du texte, il est nécessaire de ne retenir que ceux dont la

probabilité d'être de vrais mots-clés du texte est la plus forte.

Une étape suivante peut consister à filtrer les groupes nominaux. Tous les groupes nominaux n'ont pas la même capacité référentielle. Certains sont plus importants que d'autres. Pour déterminer quels sont les plus importants d'entre eux, le système selon l'invention valorise chaque groupe nominal en fonction d'un double critère, l'un statistique, l'autre syntaxique.

Le critère statistique :

Les mots les plus fréquents des groupes nominaux sont classés par ordre de fréquence décroissant (en tenant compte d'une approximation comme 'comédie' = 'comédies'), soit dans le texte d'exemple :

comédie	3
musicale	3
affiche	2
années	2
Cats	2
dernière	2
représentation	2

Seuls les mots dont l'occurrence dépasse 1 sont conservés dans la liste. Les mots éliminés ont donc une valeur nulle. On ajoute à la valeur de chaque groupe nominal (initialement fixée à 0), la valeur de l'occurrence des mots qu'il contient moins 1. La valeur des groupes nominaux devient :

comédie musicale	(3 - 1) + (3 - 1) = 4
affiche	2 - 1 = 1
affiche à Londres	2 - 1 = 1
Cats	2 - 1 = 1
etc...	

Le critère syntaxique :

Lorsque qu'un groupe nominal est ou comporte un nom propre, celui-ci prend un point de valeur supplémentaire, 0 sinon.

comédie musicale	4 + 0 = 4
affiche	1 + 0 = 1
affiche à Londres	1 + 1 = 2
Cats	1 + 1 = 2
etc…	

Avec cette valorisation, il est aisé de procéder au classement des groupes nominaux. Dans le texte d'exemple, les groupes nominaux perçus comme les plus importants sont soulignés deux fois, les groupes d'importance secondaire sont soulignés une fois, tandis que les autres ont été purement et simplement éliminés.

«Cats», l'une des comédies musicales les plus longtemps à l'affiche, va tirer sa révérence après vingt et une années sur la scène londonienne. La dernière représentation de cette œuvre d'Andrew Lloyd Webber aura lieu le 11 mai, jour de son 21e anniversaire, après quelque 9 000 représentations. L'annonce a été faite trois jours après la dernière représentation de «Starlight Express», la seconde comédie musicale la plus longtemps à l'affiche à Londres, après dix-huit années sur les planches.

La fin de «Cats» est un coup dur supplémentaire pour le quartier de Covent Garden, où sont regroupés la plupart des théâtres londoniens, et qui a souffert d'une forte baisse de fréquentation en 2001. Depuis 1981, année de son lancement, la comédie musicale a, depuis, été interprétée devant plus de 50 millions de spectateurs en 11 langues et dans 26 pays.

(source Reuter)

Revendications

1. Procédé d'extraction d'informations dans un texte en langage naturel, par identification de motifs (patterns), caractérisé en ce que l'on effectue un codage des mots du texte en les comparant avec le contenu d'un lexique prédéfini de mots outils, et en ce que l'on identifie ensuite des groupes nominaux en recherchant, parmi des sous-ensembles de la suite des mots codés ainsi obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies.

2. Procédé selon la revendication 1, caractérisé en ce que le codage des mots du texte s'effectue par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit lexique de mots outils, de façon à repérer les mots outils dans le texte et en ce que la fonction des mots d'usage, non reconnus comme mots outils, est déduite en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils.

3. Procédé selon l'une des revendications 1 ou 2, caractérisé en ce que les groupes nominaux identifiés sont ensuite valorisés de façon à ne retenir que les groupes perçus comme les plus importants en utilisant des critères de valorisation prédéfinis.

4. Système d'extraction d'informations dans un texte en langage naturel, caractérisé en ce qu'il comprend :

- une unité d'entrée pour recevoir ledit texte en langage naturel,
- un fichier lexique dans lequel sont enregistrés des mots outils,
- un processeur d'analyse relié à ladite unité d'entrée, au fichier lexique et agencé pour effectuer dans un premier temps le codage des mots dudit texte en langage naturel par évaluation de la fonction grammaticale de chaque mot en le comparant avec le contenu dudit fichier lexique de mots outils, de façon

d'une part à repérer les mots outils dans le texte et à évaluer la fonction des mots d'usage, non reconnus comme mots outils, en comparant leur emplacement par rapport à l'emplacement des mots reconnus comme mots outils, et dans un deuxième temps une recherche, parmi des sous-ensembles de la suite de mots codés obtenue, des groupes de mots codés répondant à des règles syntaxiques prédéfinies, de façon à identifier des groupes nominaux,

- une unité de sortie reliée audit processeur d'analyse pour recevoir les groupes de mots codés reconnus comme des motifs syntaxiques.

5. Système selon la revendication 4, caractérisé en que le processeur d'analyse comprend en outre des moyens de valorisation des groupes de mots codés retenus de façon à ne retenir que les groupes perçus comme les plus importants.

6. Système selon l'une des revendications 3 ou 4, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de reconnaissance de la langue du texte reçu dans l'unité d'entrée.

7. Système selon l'une des revendications 4 à 6, caractérisé en ce que le processeur d'analyse comprend en outre des moyens de régularisation du texte reçu dans l'unité d'entrée de façon à éliminer les amalgames de signes.

Fig. 1

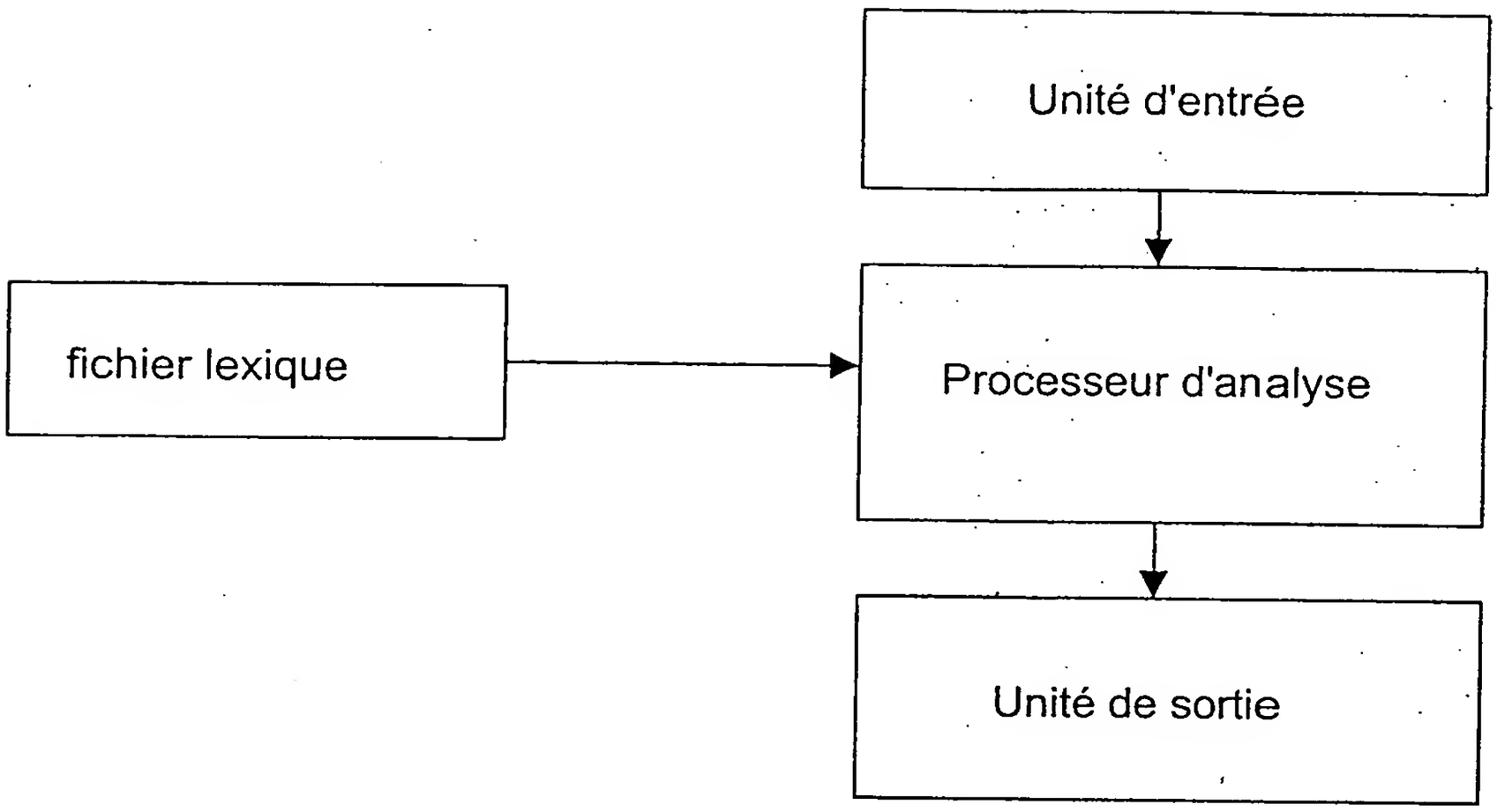
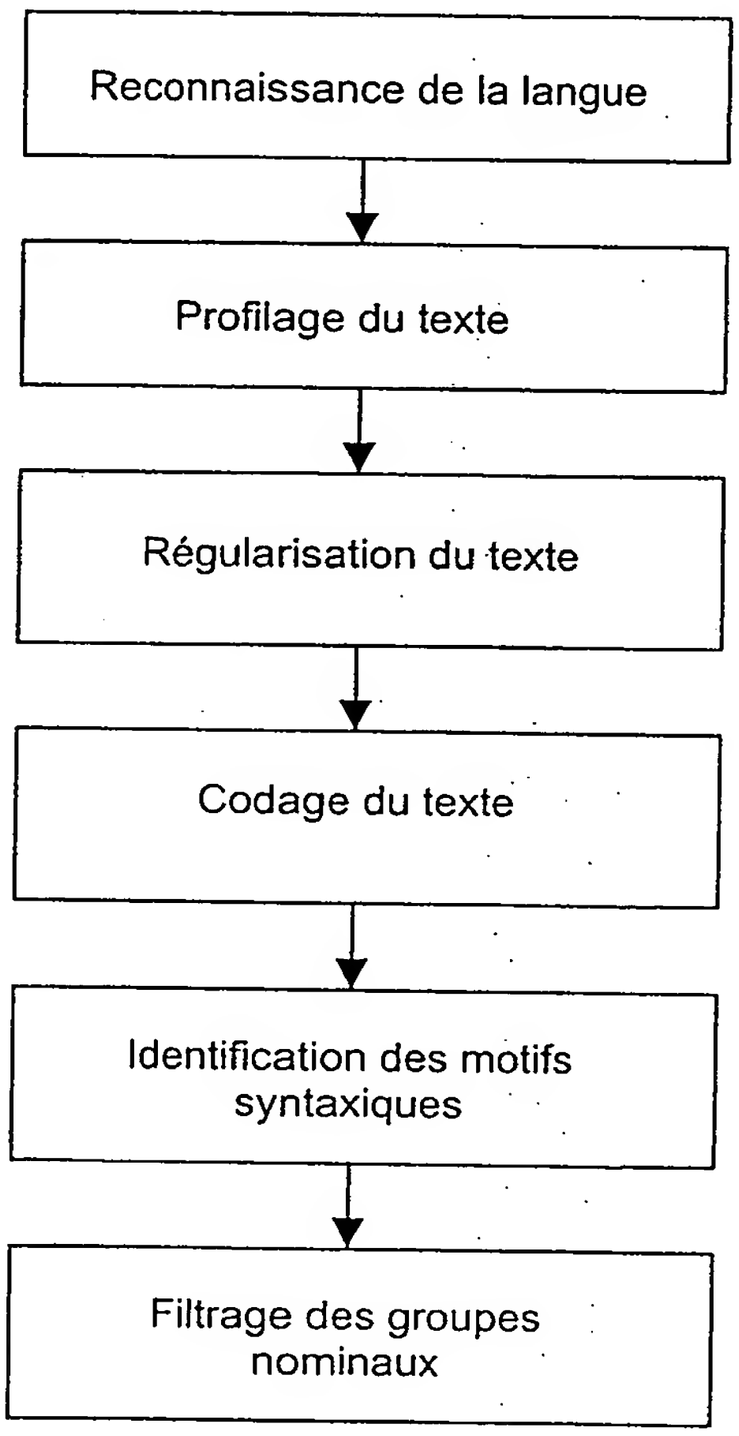

Unité d'entrée
fichier lexique
Processeur d'analyse
Unité de sortie

Fig. 2


Reconnaissance de la langue
Profilage du texte
Régularisation du texte
Codage du texte
Identification des motifs syntaxiques
Filtrage des groupes nominaux